Computational design of orthogonal microRNAs for synthetic biology

Maciej Trybiło

School of Information Systems, Computing and Mathematics

Brunel University

A thesis submitted for the degree of

Doctor of Philosophy

June, 2013

Abstract

Upcoming applications of synthetic biology will require access to a wide array of robust genetic components (parts). The logic of a genetic system is encoded with regulatory elements such as pairs of transcription factors:promoters, miRNAs:target sites, or ribozymes:aptamers among others. Due to a relatively simple form and mode of operation of miRNAs, it is possible to design their synthetic variants. Out of all possible miRNA sequences the ones chosen should perform efficiently and should avoid cross-talk with both the host system circuits and within the imported synthetic ones. In this work, a computational method involving a series of heuristics is developed that can be used to design ensembles of such sequences depending on the host transcriptome. As an example, an ensemble of eight such miRNA sequences is produced using this method for use in a human host. Those have then been validated experimentally against the above-mentioned requirements by transfection into HEK 293 cells and flow cytometry measurements of fluorescent markers. The produced sequences are available for use from pENTR vectors of the Gateway cloning system. The required computations were facilitated by a modern cluster computing system—Kaichu—especially developed for this project, but fit for general purpose use and available under an open-source license.

Contents

List of Figures

Acknowledgements

This work would not be possible without the people around me.

First, I would like to thank my supervisors. Prof. David Gilbert, and Dr Amanda Harvey always strived to provide me with the best guidance, work environment, funding and training possible. They have ensured that I would be exposed to the best science at important conferences, seminars and other events. I would also like to thank them for their patience and encouragement, especially during the write up.

Special thanks go to my collaborator Dr Liliana Wróblewska from the Ron Weiss lab at the Massachusetts Institute of Technology who generously proposed to help me with the experimental work. Her explanations and example taught me a lot about experimental work and her enthusiasm motivated me a great deal. Dr Wróblewska has put a large amount of time, effort and work into helping me with the experimental design and performing a lot of the experimental work herself.

I would like to thank Prof. Ron Weiss himself for supporting this collaboration, having me as a visitor in his lab, and allowing the MIT part of the experiments to be funded there.

My parents—Krystyna and Wiesław—encouraged me and supported financially when it was needed the most.

I would like to thank the members of the Gilbert lab for being kind and helpful cohabitants of the attic office: Pam Gao, Crina Grosan, and Zujian Wu.

During the study I was lucky to make true friends who made the lonely doctoral endeavour wholesome. Big thanks to all of the current

or soon-to-be PhDs: Amirahmad Bigdeli, Stefano Ceccon, Chandrika Cycil, Xavier Duportet (MIT), Bahareh Heravi, Panagiotis Panagiotopoulos, Sara Robaty, and Fotis Spyridonis.

Thank you to Haliun Altankhuu who always was there through good and bad.

I would also like to thank my co-workers at We7/blinkbox music who were rooting for me during the write up. Special thanks there go to Dr Sven Schmidt and Duncan Gossage for proofreading.

I am very grateful to the EPSRC who provided me with full funding for three years of the project.

Author's declaration

I hereby declare that I am the sole author of this thesis.

I authorise Brunel University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Signature

Date:

I further authorise Brunel University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature

Date:

Chapter 1

Introduction

Synthetic biology is a rapidly developing new approach to engineering biological systems that aims to utilise engineering principles such as modularity, reusability, and standardisation in order to efficiently manage the immense complexity of its problem domain. The applications of synthetic biology span from synthesis of complex chemicals, biomanufacturing, sensing, to medicine. A biological system is in principle capable of efficient production of biofuels, drugs, and other high-value compounds. New materials could be produced with atomically-precise manufacturing. Engineered bacteria could sense and alert to dangerous chemicals in the environment. One could envisage a modified bacterium that, once injected into the bloodstream, would evade the immune system, and specifically recognise and kill cancer cells.

This great promise is however yet to be realised and major efforts are required to deal with the many teething problems that stem from our lack of knowledge of the natural biology, the incredible complexity and the design style (or lack thereof) of the living systems, and the inherent stochastic noise of chemical interactions.

The behaviour of biological systems emerges from the information encoded in the DNA. This information plays out on different levels of control from gene expression regulation through protein-protein interaction networks, hormonal regulation to the nervous system. In this work the natural control mechanisms are seen from the point of view of their utility in biological engineering. The focus of synthetic biology today is on the lowest level of biological control which is

the regulation of gene expression. Here, the genetic program states how expression of particular genes, modulated by signals from the environment, affects the expression of other genes.

Different mechanisms of genetic regulation exist. The one that is understood best and is most widely used in synthetic biology is transcriptional regulation by promoters and transcription factors. In this mechanism a sequence upstream from a regulated region (the promoter) is specifically recognised by a protein (transcription factor) in order to drive or inhibit the transcription of the regulated region. This widely used mechanism is not without its limitations. Due to the nature of the interaction between the transcription factor and the promoter it is difficult to engineer the regulatory pairs. This means that the choice of different regulatory parts of this kind with different characteristics will be limited. The model that transcription is carried out at a smooth and steady rate is untenable as transcription is particularly subject to stochastic noise and is carried out in bursts by the cell. These phenomena often cause synthetic circuits to behave in unexpected ways.

But nature can sometimes be very precise. The roundworm *Caenorhabditis elegans* in its adult hermaphrodite form always comprises 959 cells (Sulston and Horvitz, 1977) out of which exactly 302 are neurons, always connected into the same network (Hobert, 2005). During the development exactly 131 cells undergo apoptosis (programmed cell death) (Sulston and Horvitz, 1977). The whole lineage of cells from zygote to the adult form is known and essentially fixed. Such precision requires very robust control.

In most eukaryotes, a different kind of genetic regulation that has particularly beneficial properties is translational regulation provided by the RNA interference pathway. In this mechanism, short ($\sim$22 nt) RNA molecules (the interfering RNA) can recognise a subsequence (a target site) of an mRNA transcript and inhibit its translation into protein. The recognition of a target site by an interfering RNA depends on partial complementarity between them. Those interfering RNAs are differently named depending on their source, but the mechanism of regulation they are involved in is the same. As will be explained in Chapter 2, the most useful from the point of view of synthetic biology are microRNAs (miRNAs).

RNA interference is a flexible mechanism as the target site can be placed in

the untranslated regions of the mRNA and thus any protein can be put under control of any interfering RNA. Furthermore, it is possible to put one transcript under the regulation of more than one interfering RNA. The latter property is in contrast with the typical transcriptional regulation.

Since RNA interference works at the translational level, it can be used in tandem with transcriptional regulation. The central idea here is that miRNAs can be used to sharpen the regulatory response and fine tune protein levels. In fact, this is how they are used in nature. miRNAs have been shown to be crucial in the development of many species, including *C. elegans* (Miska et al., 2007; Reinhart et al., 2000; Wightman et al., 1993).

Considering the properties of the RNA interference pathway and that its proof of concept has already been provided to us by nature it is becoming clear that this will be a staple mechanism in future synthetic biology. For someone who wants to introduce miRNAs into their synthetic circuits it would be obvious to take the same route that had been taken with regards to the transcription factors and promoters—use the pairs that already exist in nature. More can be done however as is explained below.

The nature of the interaction between the interfering RNA and its target site as well as the simple linear structure of both of the regulatory elements should mean that it is relatively easy to design completely new regulatory pairs. It should also be possible to make those pairs orthogonal with respect to the rest of the system.

The issue of orthogonality is one of the most important in synthetic biology. Unlike in, for instance, electronic engineering there is no or little physical separation between signals in the cell. Transcription factors and miRNAs move more or less freely inside the nucleus or the cytosol in the cell and can target any matching gene or mRNA transcript respectively. In this case, for signals to be separate, they have to be incompatible with each other. Thus, to avoid cross-talk between parts of the system, each regulatory pair can essentially be used only once and has to be sufficiently different from all of the other ones. This exclusivity also regards the regulatory pairs existing in the entirety of the indigenous system as well.

Considering the above the motivation of this work is to provide means to

design new regulatory miRNA:target site pairs that can be used in synthetic circuits and will fulfil the orthogonality requirement.

This work is about the design and implementation of a software system that produces such regulatory pairs. Furthermore, a library of eight miRNA:target site regulatory pairs has been calculated using the software for a specific host—the human cells. Those pairs have then been tested in the wet-lab to check for the inhibition strength and for orthogonality between each other.

In the course of this work a new general purpose Master-Worker cluster computing framework—Kaichu—has been developed and has been released to the community under an open source licence.

1.1 Contributions

MicroRNAs are robust regulatory elements found in all multicellular organisms, performing crucial functions. Better availability of new, efficient , and orthogonal miRNAs will lend itself to progress in synthetic biology. As part of this project the following contributions to knowledge have been made.

1.1.1 A framework for design of novel miRNA sequences

A software framework has been developed that can be used to rationally design novel miRNA sequences. The framework attempts to ensure that the discovered sequences are orthogonal with respect to a specified transcriptome and they are provided in ensembles where the sequences are orthogonal with respect of each other. For the reasons of orthogonality with respect to the transcriptome those ensembles are host-specific and the possibility of reuse of the framework is crucial for future benefit. The framework follows a workflow consisting of algorithms adapted or designed for their purpose.

1.1.2 A library of 8 synthetic miRNAs

The mentioned framework has been used to produce an ensemble of eight synthetic miRNA sequences for specific use in human cell lines. The miRNA expres-

sion units and target sites corresponding to those sequences have been subsequently synthesised and cloned into DNA plasmids, which makes them readily available for the synthetic biology community. The ensemble has been tested in HEK 293 cells to confirm that the miRNA:target site pairs perform well and that there is no cross-talk between the pairs.

1.1.3 A cluster computing framework—Kaichu

Relatively easy availability of spare computational power from resources such as teaching machines at universities motivates researchers to develop their software to run in a parallelised fashion. Such resources can be utilised outside of the working hours, so the machines are fully available to the computations, but that still leaves a large potential untapped. On the other hand developing distributed systems that are correct and efficient in an uncertain environment is difficult, and the available solutions are unwieldy for small projects or fast-changing code bases.

A cluster computing framework called Kaichu has been developed to aid the parallelisation of the computations required in the main work in this project, but designed for general use. In this project, Kaichu performed in an uncertain environment such as a pool of computers that often become available and unavailable. It has a number of unique features, such as automatic job aggregation. The two main design goals were to enable rapid development of parallelised software and efficiency.

1.2 Structure of the thesis

Synthetic biology marries methods and approaches from biology, computing, and engineering. Such an interdisciplinary field is a meeting of different cultures in terms of language and methods, and also the shape of a typical doctoral thesis.

This work has motivation rooted in biology, has a computational solution, and is concluded with biological verification. The three parts are written in respective styles to ensure reproducibility of the work. However, given that the thesis is brought forth in a department of computing, the biological parts assume

computational audience with little background knowledge.

Chapter 2 - Background provides a more detailed view of synthetic biology. The biology of the RNA interference pathway and of miRNAs in particular is described including their biogenesis, current understanding of the rules that govern the targeting mechanism, and some examples of miRNA function in natural circuits. The engineering point of view of the RNA interference pathway is presented to describe its potential in regulatory circuits. Finally, examples of miRNAs being used in synthetic biology so far are discussed.

Chapter 3 - Design and Implementation of Software spells out the exact requirements for the software that designs libraries of novel miRNAs. The orthogonality requirement is analysed into two orthogonality requirements—orthogonality with respect of the host transcriptome and intra-library orthogonality. The computational challenges that stem from those requirements are considered. Then it follows up with the design decisions that have been made to meet the requirements. The heuristic strategies and the composition of optimisation and graph algorithms has been developed in order to make the computations tractable.

Chapter 4 - Kaichu Cluster Computing Framework presents the general purpose Master-Worker cluster computing framework that has been developed in the course of this project. Its focus is on enabling rapid development of distributed computation applications that are efficient and dependable in uncertain environments such as student lab machines in educational institutions.

Chapter 5 - Materials and Methods gathers detailed protocols of the wet-lab work performed as part of this project. Such description is necessary to ensure reproducibility of the work.

Chapter 6 - Experimental Design describes the design of the experiments that have been carried out in order to test for the efficiency of the new microRNAs and their inter-ensemble orthogonality. A test construct has been developed wherein the miRNAs were co-expressed with a red fluorescent protein, and a cyan

protein was put under the control of respective miRNAs. The repression strength was measured using flow cytometry as percent decrease in cyan fluorescence.

Chapter 7 - Results reports and discusses the results of the experiments set out in Chapter 6. The chapter starts with a history of an early setback and then proceeds to present the results that validate the final set of the synthetic miRNAs in requirements of efficiency and inter-ensemble orthogonality.

Chapter 8 - Conclusion provides a conclusion to the entire work. The research questions are restated and motivated. The empirical findings are recounted and a more in-depth analysis is provided. Applications of the work are discussed and possibilities for further work—improvements, and other future avenues of research—are considered.

Chapter 2

Background

2.1 Synthetic Biology

The definition of synthetic biology is surrounded with some controversy. In this work synthetic biology is understood as a discipline that aims to manipulate and build biological systems following engineering principles and techniques such as standardisation, modularisation and modelling. Most useful or interesting engineered systems will carry complexity that has to be managed.

Standardisation is pervasive throughout synthetic biology. There are attempts to standardise lab procedures to ensure repeatability (Kelly et al., 2009). Characteristics of DNA parts are being standardised to ease comparisons between them. For instance, promotors are characterised with PoPS (Polymerase per second) to measure the strength of transcription signal, however what constitutes a characterisation of a part or what constitutes a part is still far from being established. Lastly, DNA assembly methods are being standardised.

Modularisation is a technique of separating a system into modules that have defined behaviours and interfaces. Such separation of concerns simplifies analysis of a designed system and fosters reusability. In other engineering disciplines, such as software engineering or electronics, modularisation is easily achieved technically because it is easy to separate signals. For instance, in electronics electrical signals are routed through wires of a conductor. In the context of a cell, with some simplification, all signals are mixed together. A concentration of a tran-

scription factor will affect all modules that accept it as their input. This means that modularisation has to be achieved in novel ways in synthetic biology.

Complex systems can be thoroughly understood through creation and analysis of their models. Models can also serve as a specification and document the system as well as allow to explicitly state the assumptions under which the system is designed. It is common to represent biological systems as diagrams, but those informal models are not sufficiently accurate to be really useful. Formalised models include time-free qualitative, such as qualitative Petri nets where only the structure of a biochemical network is represented - the species involved and possible interactions between them. Quantitative models add time and aim to capture the dynamic behaviour of a system. Those can be distinguished into continuous models that are based on ordinary or partial differential equations (ODEs and PDEs) and stochastic models which deal with discrete entities and events that happen with certain probabilities (Gilbert et al., 2008). The stochastic models aim to capture the stochastic noise that manifests itself in chemical reactions with relatively few entities present.

2.1.1 Applications

Living systems can manipulate matter and energy in an array of ways, they are also capable of processing information at a molecular level. The applications of synthetic biology are wide ranging and to a large extend still unforeseen. It is possible however to list some of the areas where impact of synthetic biology can be anticipated to be the greatest. Pragmatically, it is important to consider potential applications in the context of current technology, especially electronics.

Living cells are complex chemical factories whose capabilities go far beyond the achievements of modern synthetic chemistry. Metabolic engineering can enable us to produce **high-value chemicals** (pharmaceuticals, etc) at low cost and synthetic biology can be used to build and optimise anabolic networks to achieve that. For instance, a group at UC Berkeley managed to engineer *E. coli* to efficiently synthesise a late precursor to artemisinin - an effective antimalarial agent (Wu et al., 2011). One of the most commonly mentioned biosynthesis applications is production of biofuels. This is however controversial as it would

often mean competition for large areas of arable land with food crops and might never be as efficient as photovoltaics in turing sunlight into usable forms of energy (Blankenship et al., 2011).

Advanced genetic manipulation opens possibilities for radically novel **medical therapies**. One could imagine engineered viruses or bacteria injected into the bloodstream and performing medical interventions in the body. For instance bacteria could recognise and kill cancerous cells. Xie et al. (2011) have already managed to construct a circuit that specifically triggers apoptosis in HeLa cells by their miRNA profile. Synthetic circuits can also enable new methods for drug discovery.

An important requirement to survival of a living system is to be able to sense the state of the environment. For this reason living organisms have been endowed with a wide range of sensory capabilities. Cells alone can detect chemical compounds, temperature, electromagnetic radiation of different wavelengths, mechanical pressure, or electric current. Detection of chemical compounds is the most difficult given the current state of technology, and that is where synthetic biology could make the highest impact in the context of **biosensing**. Organisms harbouring synthetic constructs could act as transducers from a chemical signal to an easily detectable signal such as electric current (Gu et al., 2010), or fluorescence (de Mora et al., 2011). Typical areas could include detection of pollutants de Mora et al. (2011); Gu et al. (2010) or medical diagnosis (Saeidi et al., 2011).

Other applications include **biofabrication** of materials such as spider silk (Widmaier et al., 2009), or stimulation constructs that allow precise **studies of molecular biology**.

The abovementioned application areas are those considered most often. However, given the capabilities of biological systems those applications might prove very restrictive in the future. One example of a creative use of biological systems is work by Kemmer et al. (2011) on optimised cow fertilisation. The authors have engineered a construct that would recognise ovulation and degrade microcapsules, releasing sperm in an ideal time for fertilisation, removing the need for trial and error.

2.1.2 Challenges in synthetic biology

Synthetic biology as a field has been making a steady albeit slow progress in the past few years. There are few accomplished applications and most of the achievements were in building small synthetic modules that should become the building blocks of the future systems. Such modules are often digital and analog control and memory elements inspired by electronics: switches (Deans et al., 2007; Gardner et al., 2000; Kramer et al., 2004), logic evaluators Rinaudo et al. (2007*a*); Win and Smolke (2008*a*), counters (Friedland et al., 2009), filters Basu et al. (2005); Sohka et al. (2009), oscillators Elowitz and Leibler (2000); Fung et al. (2005); Stricker et al. (2008); Tigges et al. (2009).

The apparent inability to go beyond simple construct proves that biology is hard to engineer. A review by Purnick and Weiss (2009) showed that the complexity of published synthetic circuits, as measured by the number of promoters used, had quickly plateaued to about six. The authors identify the achievements in synthetic biology so far as the first wave. They expect that in the second wave "synthetic biologists will formulate new and effective bioengineering design principles to address these challenges." Those improvements have yet to come to fruition.

Most of the difficulties stem from the fact that our understanding of biology is still severely lacking. Not only do we often not fully understand the basic processes such as transcriptional or translational regulation, but we struggle discovering principles upon which natural biological systems are built. Given that life has been shaped by evolution rather than a rational process this is a huge challenge. Evolution sometimes seems to produce elegant, modular designs and sometimes interference between parts of the natural systems makes the workings of it seemingly impossible to untangle and, more importantly, rationally modify.

Molecular processes are subject to stochastic noise. Such effects can easily deregulate an idealistic circuit. Furthermore, behaviour of synthetic circuits is highly context-dependent. The same circuit that works as expected in an *E. coli* is likely to fail in *B. subtilis*. Synthetic circuits from the "second wave" should be able to adapt to such variations for instance by featuring stabilising negative feedback loops. In some situations stochastic noise could even be embraced and

used to generate probabilistic effects. For instance, it is known that stochastic noise in protein expression are important for determining differentiation fates (MacArthur et al., 2009).

The number of different genetic parts that are reasonably well understood and used is small. Most of the published designs share commonly used elements such as the parts of the *lac* operon. Due to lack of separation of signals in the cell such modules cannot be used together in one system as they would interfere with each other. A significant effort is needed to characterise more parts and understand interactions between them.

Engineering disciplines are typically heavily aided by software - CAD (computer-aided design) tools help design and model complex structures from microprocessors to skyscrapers, CAM (computer-aided manufacturing) aid the production. In synthetic biology such software is still in its infancy. The whole chain of development and analysis of the abstract design, instantiation into concrete parts, to DNA sequence, DNA assembly, to testing should be captured in a software-aided process. Most modelling software such as CellDesigner (Funahashi et al., 2003), COPASI (Hoops et al., 2006), SimBiology (Liu et al., 2010) has been developed with systems biology in mind and, however useful, does not make provisions for construction of synthetic systems. Software that has ambitions to be CAD for synthetic biology is starting to appear with examples such as TinkerCell (Chandran et al., 2009) or GenoCAD (Czar et al., 2009).

The challenges in constructing reliable synthetic circuits will inadvertently require a lot of trial and error, if only to aid our understanding of biological principles for the future. It is often impossible to predict which exact DNA sequence will result in the desired behaviour and many variations might need to be tested. This means that the cost of building and testing should decrease both in terms of time and resources. The cost of DNA synthesis is decreasing, but it is still not viable to synthesise whole plasmids from scratch. Instead, error prone and time consuming DNA assembly and cloning has to be used. Such standard laboratory methods should be automatised with robotics to allow reproducibility and more information feedback to the designer.

2.2 RNA design in synthetic biology

RNA has been recognised to provide much wider ranging functionality than to be just a passive temporary store of information such as is mRNA. In the context of synthetic biology, the most interesting is the capability of RNA to control gene expression both at the transcriptional and translational level. The relative predictability of RNA hybridisation has been exploited before to design novel sequences that can be used in those mechanisms in the synthetic biology context.

In bacteria, translation can be controlled by conditionally blocking the ribosome binding site (RBS). NOT gates can be made when a short RNA (sRNA, a riboregulator) can hybridise to and around the RBS of a specific mRNA, thus preventing translation. YES gates can be constructed by making the mRNA folding in a way to sequester the RBS and a specific short RNA can bind to the RNA such that it unfolds the mRNA and exposes the RBS again.

In this mechanism the required general structure of the involved species is known and an example sequence fulfilling this structure can be easily devised, however it has to be ensured that the dynamic behaviour of transformations between the folded and hybridised states is performed. This is based on the balance of free energies of all of the states of the system—the opposing requirements of stability of each configuration and propensity to transform between the configurations have to be satisfied. Rodrigo et al. (2012) devised computational methods that automatically design sequences that fulfil those requirements. In Rodrigo et al. (2012), a Monte Carlo simulated annealing is employed to solve the inverse folding problem to design the mRNA and a controlling sRNA sequences that perform as a YES gate. The algorithm starts from example sequences that satisfy the required secondary structures in both configurations. Then those sequences are mutated while retaining those structures within a certain margin and optimise the objective function. The objective function takes into account the free energy of the stabilised folded structures and the activation energy between this configuration and the transition state when the two species start to interact and hybridise. The resulting candidates were then confirmed to perform as intended *in vivo* and an AND logic gate was successfully constructed using allosteric transcriptional regulation together with an sRNA YES gate.

Rodrigo et al. (2013) generalise the approach taken in Rodrigo et al. (2012) and expand the heuristic method to design higher order systems. For example, a circuit was designed where a single sRNA specie was involved in two different gates (a NOT and an AND gate).

Win and Smolke (2007) combined hammerhead ribozymes and ligand-controlled aptamers to implement riboswitches in bacteria. A hammerhead ribozyme is a cis-acting RNA controller sequence that folds into a three-loop secondary structure and causes rapid degradation of an mRNA it is incorporated into. An aptamer is an RNA sequence that can specifically alter its secondary structure based on interaction with a small molecule (ligand). Such aptamers can be selected for *in vitro* to bind to a chosen ligand and be highly specific (Ellington and Szostak, 1990; Tuerk and Gold, 1990). An aptamer can be incorporated into one of the side loops of a ribozyme and switch it in and out of its active state. Such system can be integrated into the 3' UTR and thus translation of an arbitrary gene can be put under control of a small molecule.

The authors took advantage of two mechanisms of signal transmission from the aptamer to the ribozyme: strand displacement and helix slipping. In the latter case, the design of functional systems was based on *in vivo* screening. The design of the systems based on strand displacement was based on rational strategies that take into account the balance of free energies between secondary structure conformations of the switch. The authors supported their rational strategies with secondary structure prediction and free energy calculation software, but did not report synthesising those strategies into a computational framework.

The riboswitch platform was later made more modular by incorporating a third element—a transmitter—between the ribozyme and the aptamer that would be solely responsible for transmitting the regulatory signal from the aptamer and the ribozyme (Win and Smolke, 2008*b*). This also allowed the aptamers to be chained. Furthermore, the researchers were able to build AND, NOR, NAND, and OR gates by incorporating aptamers in both of the side loops of the ribozymes or by integrating more than one switch into a single 3' UTR.

2.3 MicroRNAs and siRNAs in nature

MicroiRNAs and siRNAs are very closely related and differ only in their biogenesis while having the same mode of operation. While siRNAs are produced from long stretches of double stranded RNA that are either transcribed endogenously or have an external source, miRNAs are produced from folded single stranded RNA that is transcribed endogenously and are primary components of the gene regulatory system. From the synthetic biology perspective miRNAs are more interesting as the applications will require their established presence and expression rather than depend on administration of siRNAs. However, much of the research into the mechanisms of RNAi has been done with focus on siRNAs for their usefulness in genetic studies. For instance, the siRNA design rules have been established for that purpose. Those findings usually also apply to miRNAs and siRNAs are considered throughout this chapter.

The first hints of sequence-specific, RNA-induced gene silencing appeared in the 1980s when Izant et al. (1984) discovered that antisense RNA can inhibit production of specific genes. This phenomenon was later studied more and exploited (Fire et al., 1991; Guo et al., 1995; Nellen and Lichtenstein, 1993).

However, a breakthrough in understanding of the RNA interference (RNAi) was brought by Fire et al. (1998) who, working in *C. elegans*, discovered that it is the double-stranded RNA (dsRNA) that is really the causative agent of this mechanism. The authors provided some evidence that RNAi happens post-transcriptionally as targeting promoter regions and introns did not result in silencing as well as postulated an additional catalytic element and amplification of response, both of which were later validated.

Andrew Fire and Craig C. Mello were later awarded the Nobel Prize in Physiology and Medicine for this work.

Hamilton and Baulcombe (1999) found out that post-transcriptional gene silencing (PTGS) - another name for RNAi - is mediated by RNAs of 25 nt length. Although they thought that those were single-stranded, they were the first to point to the short length of the agents of RNAi. Those that are produced from longer stretches of dsRNA are called short interfering RNAs (siRNAs).

The Tuschl Lab was the first to confirm that siRNAs also work in mammalian

cells by co-transfecting 21 nt long siRNA duplexes with reporter constructs (Elbashir, Harborth, Lendeckel, Yalcin, Weber and Tuschl, 2001). The same lab provided evidence that 21 and 22 nt long siRNAs are produced by cleaving dsRNA in an "RNase III-like mechanism" leaving the duplexes with 2 nt 3' overhangs, and that the cleavage of the target RNA happens near the centre of the target site (Elbashir, Lendeckel and Tuschl, 2001). Later they also established that the best-performing duplexes have 2 nt long 3' overhangs, and are 21 nt long (Elbashir, Martinez, Patkaniowska, Lendeckel and Tuschl, 2001). Furthermore, they have found out that the efficiency of knockdown was affected more if mismatches between the antisense strand and the targeted mRNA were introduced on the 3' side of the antisense strand. siRNAs were considered a tool for molecular biology study that allowed selective gene knock down and knock out, so the authors took a practical approach and started a trend to discover siRNA design guidelines to help researchers design efficient siRNAs.

To summarise the above, microRNAs (miRNAs) are short (∼22 nt), non-coding RNAs that, as part of the RNAi (RNA interference) pathway, are involved in translational regulation of most eukaryotic organisms (Ambros, 2003; Bartel, 2004; Lai, 2003). MicroRNAs hybridise to mRNA transcripts and repress their translation by either causing their degradation or otherwise interfering with ribosomal action. The evolutionary origins of RNAi are thought to lay in prokaryotic anti-viral defense and RNA processing. As a complementary genetic regulation mechanism to transcriptional regulation, miRNAs are involved in important control processes in development and physiology.

The miR-143 and miR-145 regulate smooth muscle proliferation and differentiation in mouse (Cordes et al., 2009). The existence of miRNAs goes towards explaining the complexity of higher organisms in spite of their relatively low number of protein coding genes. MiRNAs are known to play a role in neuronal development in *C. elegans* (Hobert, 2006) as well as mouse and human (Lagos-Quintana et al., 2003; Sempere et al., 2004).

van Rooij et al. (2007) discovered a negative feedback loop involving miR-208 in the heart muscle. In this circuit miR-208 is expressed together with the αMHC which is induced by the TRAP complex. The miRNA in turn represses the production of the THRAP1 component of TRAP to stabilise the action of the

complex. Overexpression or absence has been shown to deregulate the activity of THRAP1 and negative phenotypic effects.

Analysis of natural regulatory networks suggests that miRNAs are used both in sharpening the expression profile and fine-tuning of gene expression (Stefani and Slack, 2008). This is possible because miRNAs operate in a wide range of efficiency depending on the level of complementarity between the miRNA and their target.

MiRNAs are not limited to animals or multicellular organisms. They are also found in single cellular algae (Molnár et al., 2007) as well as in some species of budding yeast (Drinnenberg et al., 2009).

2.4 MicroRNA biogenesis

The production of miRNAs is carried out in several steps (Figure 2.1). MiRNAs are transcribed by Polymerase II from independent transcription units, polycistronic clusters or introns of protein-coding genes (Bartel, 2004). Typically, the initial transcript (pri-miRNA) is in the order of hundreds of nucleotides and folds into a structure with many hairpins. The Drosha enzyme cleaves the pri-miRNA to release ∼65 nt long hairpins with a 2 nt overhang at their 3' end (pre-miRNA). The pre-miRNA is exported into the cytoplasm by exportin-5 where the hairpin loop is excised by Dicer to leave a duplex with a 2 nt 3' overhang on each side. One of the strands is preferentially chosen to be incorporated into the RNA-induced silencing complex (RISC) and becomes functional (the guide strand) while the other is degraded (the passenger strand). The rules for which of the strands is chosen to be the guide strand are captured under the siRNA design rules (Section 2.6). It is thought that it depends on the relative thermodynamic stability of the duplex and that it should be lower at the 5' end side of the guide strand relative to the other end (Reynolds et al., 2004; Ui-Tei et al., 2004).

2.5 MicroRNA mechanism of operation

The miRISC binds to mRNA transcripts based on full or partial complementarity with the miRNA. There are two modes in which the actual repression happens.

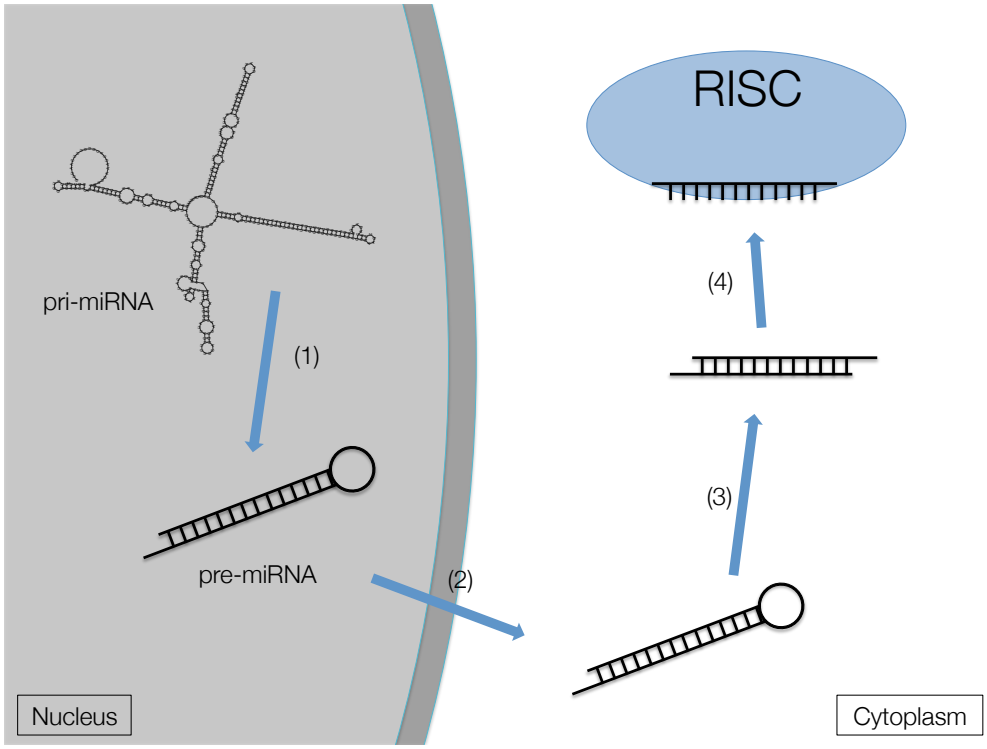

Figure 2.1: **Schematic of miRNA biogenesis. (1)** The pri-miRNA is cleaved by Drosha to release the pre-miRNA hairpin. **(2)** The pre-miRNA is exported into the cytoplasm by Exportin-5. **(3)** The hairpin is cleaved by Dicer to release a duplex with 2 nt overhangs on 3' ends. **(4)** One of the strands of the duplex is incorporated into the RNA-induced silencing complex (RISC) and the other one is discarded.

Full or very high complementarity results in mRNA cleavage which in turn causes its quick degradation (Pillai et al., 2007). Lower levels of complementarity result in the RISC interfering with the ribosome as it is translating the mRNA. The details of the latter are still contested.

Pivotal to this work is understanding what patterns of partial complementarity result in a functional miRNA:mRNA pair.

The early understanding of the mechanism of operation and rules that constitute a working miRNA:target site pair was that:

1. Animal target sites reside in the 3' UTR of the transcripts. This as a consequence of the location of the first discovered miRNA targets (Lee et al., 1993; Reinhart et al., 2000).

2. The seed rule that states that six (Lai, 2002; Lewis et al., 2005) or seven nucleotides, starting from the second position from the 5' of the miRNA have to be fully base-paired to the target site.

Brodersen and Olivier (2009) evaluated the above in light of new evidence and concluded that:

1. The target sites are not restricted to the 3' UTR and any exposed subsequence of the transcript is a potential target site.

2. The seed is a 7 nt long region starting from the second position from the 5' of the miRNA and is more sensitive to complementarity, but does not require it in full. As a rule of thumb it can be accepted that at least five nucleotides have to keep complementarity for the pair to be functional. Additionally, up to five mismatches are accepted outside of the seed region, however not more than two consecutive ones.

It is established that the downregulation of target transcripts by small RNAs is affected by the transcript context. Abundance of transcripts that also contain appropriate target sites dilutes the effect of the small RNA. Arvey et al. (2010) tested this hypothesis and have found a significant anti-correlation between target site abundance and the strength of transcript downregulation by miRNAs and siRNAs. The anti-correlation between the target site abundance and the protein

levels downregulation followed the same trend, but lacked statistical significance. The authors used full seed matches to determine the abundance of target sites against RNA-Seq reads. RNA-Seq reads provide a better quantification of target sites than the transcriptome alone, however the authors concluded that the abundance of target sites calculated in both cases is highly correlated.

The above hypothesis is also supported by Ebert et al. (2007) who proposed using transcripts with multiple target sites as sponges (or decoys) for specific miRNAs to de-repress their original targets.

2.6 siRNA design rules

It has been discovered that sequence of an siRNA alone can bear heavily on its efficiency. This is mostly due to preferential strand selection during biogenesis. It has also been hypothesised that the short RNAs could tightly fold preventing their use as single stranded elements. Reynolds et al. (2004), based on repression experiments against 180 different siRNAs, identified eight characteristics that influenced their efficiency. According to Reynolds et al. (2004) siRNAs should have relatively low G/C content, lower internal stability towards the 5' end of the guide strand (to induce the choice of it as the guide strand), should lack inverted repeats, and comply with several additional rules for preference of certain nucleotides at certain positions. This set of rules has been widely accepted since.

However, Tafer et al. (2008) later had access to a much larger dataset of 2,182 tested siRNA sequences. Instead of eliciting concrete rules, the authors trained an artificial neural network to predict repression efficiency and managed to achieve 0.66 Pearson correlation between the predictions and the results. McQuisten and Peek (2009) took the same data and compared three machine learning methods in their ability to learn the design rules—artificial neural networks, general linear models, and support vector machines. A support vector machine achieved the best performance with 0.711 Pearson correlation.

Although being black box rather than providing explicit rules, the latter two methods provide more accurate prediction of siRNA efficiency.

2.7 RNAi in research and therapeutics

RNAi is an efficient and specific mechanism that can be easily triggered experimentally. For that reason RNAi has been widely used in *in vitro* functional genomic studies where siRNAs are administered to knock down specific genes for loss of function analysis (Bitko and Barik, 2001) and, more recently, in whole genome screens (Boutros and Ahringer, 2008)

Some of the results point to a potential for therapeutic applications of RNAi. Application of siRNA duplexes has however a limited use as usually therapies require a longer lasting effect than the lifetime of the siRNAs Karagiannis and El-Osta (2005). In order to deal with this problem, some of the research focuses on delivery of expression systems. Delivery by viral vectors proves to be a potent method albeit still can be hazardous for the patient Reid et al. (2002). Nevertheless the research continues *in vitro* and on animal models and shows potential for therapies to viral infections, such as hepatitis (McCaffrey et al., 2003; Song, Lee, Wang, Ince, Ouyang, Min, Chen, Shankar and Lieberman, 2003; Zender et al., 2003) or HIV (Song, Lee, Dykxhoorn, Novina, Zhang, Crawford, Cerny, Sharp, Lieberman, Manjunath et al., 2003), neurodegenerative disorders such as Huntington's disease (Davidson and Paulson, 2004; Xia et al., 2002), age-related ocular diseases (Reich et al., 2003), and cancer (Holen et al., 2002; Lu et al., 2003; Sioud, 2004; Wall et al., 2003).

Particular interest has been gathered around inhibition of the oncogenic *bcr-abl* gene. Experiments on the K562 leukemia cell line showed apoptosis of the cancerous cells triggered by the siRNA-induced knockdown of *bcr-abl* and sensitisation to chemoterapeutic drugs (Scherr et al., 2003; Wohlbold et al., 2003; Zaree Mahmodabady et al., 2010).

2.8 Potential for synthetic biology

It is important to note the elegant and modular design of the RNA interference pathway. The RISC complex acts as the generic active element, and the miRNA carries the targeting specification. This contrasts with elements of transcriptional regulation where the recognition and the active elements are entangled together.

Moreover, the mapping between a matching miRNA and its target site is much simpler to understand as it is based on the digital complementarity of two nucleic acids, while the relationship between a transcription factor and a promoter is based on the 3D structures of the protein and the double stranded DNA.

Similar to promoters, miRNA target sites are independent of the coding sequence they regulate as the latter can reside in the 3' UTR. However, the *cis*-regulatory elements of the RNAi offer the additional flexibility that more than one can be placed in in one 3' UTR and used independently which allows regulation of the repression strength (more than one of the same) and, more importantly, allows building NOR gates.

It has been mentioned before that miRNAs can sharpen gene regulation, where in tandem with other forms of regulation they act to bring down protein activity to essentially nothing. miRNAs also offer a sharper temporal response than transcriptional regulation. At the point of stopping transcription there are still mRNAs available to produce more protein whereas RNAi cuts off the production downstream in the process providing a faster response.

Researchers and engineers already take advantage of the properties of RNAi in their designs.

Deans et al. (2007) used RNAi to design a synthetic circuit that allows studying precise gene knock down. The circuit puts an arbitrary gene under the control of the *lac* operator, so the transcription is repressed by the constitutively expressed LacI transcription factor. LacI can be inactivated by IPTG to release the transcription. However, this set up does not guarantee full repression. Deans et al. (2007) additionally put the gene of interest under the control of a shRNA that brings its expression to undetectable levels. In order to allow the translation of the said protein in the *on* state, the expression of the shRNA is put under the control of the TetR repressor which is itself controlled by LacI. The authors were able to achieve IPTG dose dependent repression of EGFP, diphtheria toxin, and Cre recombinase.

Deans et al. (2007) envisaged their design to be used in precise studies of single genes, but this circuit can be abstracted further to impart robustness on any leaky promoter.

Ebert et al. (2007), mentioned in Section 2.3, provided a mechanism to regu-

late miRNAs themselves which is alternative to transcriptional regulation.

Rinaudo et al. (2007*a*) explored the possibilities of using siRNAs to construct arbitrary logic gates and managed to build logic evaluators with up to five variables. This work was later taken forward to detect cancer cells by their miRNA profile (Xie et al., 2011).

2.9 Summary

This chapter introduced a new approach to engineering of biological systems—synthetic biology. Synthetic biology approaches biological systems from an engineering perspective in order to cope with their complexity. Numerous applications of genetic engineering are still to be realised as engineers face epistemological and design problems.

Further, the chapter provides an overview of the history of our developing understanding of one of the most potent gene regulation mechanism—RNAi—and its agents—siRNAs and miRNAs. Special consideration is given to their mechanisms of operation and conditions for efficiency.

Finally, current uses of the RNAi from genetic studies to future therapeutic applications are reviewed as well as unique properties of the mechanism that make it especially promising for use in novel genetic circuit designs.

Part I

Computational Design of Orthogonal microRNAs

Chapter 3

Design and Implementation of Software

3.1 Introduction

This chapter describes the requirements for, design, and implementation of the software that discovers synthetic orthogonal miRNA sequences. As part of the design, several strategies have been employed to reduce the computational load of the problem. They have been put into a workflow of steps that are realised by appropriate algorithms. Among those are Graduated Optimisation, Support Vector Machines, and graph algorithms building on the Bron-Kerbosh algorithm.

The resulting software framework can be used to generate ensembles of synthetic orthogonal miRNA sequences. This software has been used to generate eight such sequences for the human host cells. Those sequences can be readily used in new synthetic circuits.

The workflow is presented in Figure 3.1 and can be summarised as follows:

1. Rank all possible seeds by the number of matches with the transcriptome.
2. Filter the seeds by applying the siRNA design rules.
3. Select a number of seeds based on their score and pairwise difference.
4. For each selected seed:

(a) Apply the SVM-based siRNA design rules to the whole sequences generated with the fixed seed.

(b) Prescreen the candidates using the Smith-Waterman with trained parameters.

(c) Score using the state of the art RNA hybridisation algorithm.

(d) Extend the sequences by one nucleotide.

5. Select one candidate from each of the small sets of candidates surviving thus far ensuring maximal pairwise difference.

3.2 Requirements

The overarching goal of the project is to be able to produce a library of miRNA sequences that can be used in synthetic circuits. Those sequences, in order to be useful, should fulfil a set of requirements. The sequences should be efficient (3.2.1) and orthogonal with respect to the transcriptome (3.2.2) as well as with respect to each other in the library (3.2.3). From those requirements stem computational challenges due to the size of the transcriptome and the number of potential miRNA candidates.

3.2.1 Efficiency

As described in Section 2.6, micro and siRNAs can have a greatly varying efficiency of target repression. There are several factors that impact the efficiency. Firstly, miRNAs are produced in several steps of biogenesis and the efficiency of each step will affect the numbers of mature sequences and their impact. Furthermore, the correct strand has to be selected as the guide strand by the cell machinery. The efficiency furthermore depends on the preferential guide strand selection. From the mi/siRNA duplex, one of the strands will be more likely to form the RISC and be active and it is crucial to select sequences that will be selected over their duplex counterpart.

Those aspects of the RNAi biology are not fully understood, but attempts have been made to capture them in siRNA design rules which take an explicit or

black box form.

3.2.2 Orthogonality with respect to the host system

One of the biggest challenges in synthetic biology today is interference of signals between the synthetic construct and the host system (Tan et al., 2009). Introduction of a synthetic circuit may have unintended consequences for the host cell and, vice versa, such interference may disturb the operation of the synthetic circuit.

It is a problem that is not encountered in other areas of engineering. For instance, in electronics the signals are clearly separated because they pass through wires insulated from each other. Biology mostly lacks such means of physical separation and in order to achieve it the signals have to be mutually incompatible.

Since, as described in Chapter 2, miRNA targeting depends on the relationship of its sequence to the sequence of the potential target, the first orthogonality requirement means that the synthetic sequences should not have a targeting relationship with too many regions of the host transcriptome.

3.2.3 Intra-library orthogonality

Orthogonality of signals is equally important within a synthetic circuit, so it is important that an ensemble of miRNAs is produced such that they do not target each others' target sites. This will be further denoted as the second orthogonality requirement.

It is worth mentioning that within a synthetic circuit built there still could be transcripts accidentally targeted by an miRNA because miRNA can hybridise at any position in a transcript. This is however unlikely and can be taken into account only when the sequence of the circuit is known.

3.2.4 Challenges

In order to explain the computational challenges in this project, first a naive solution is introduced together with its computational complexity.

3.2.4.1 Orthogonality score

The requirements of orthogonality boil down to the relationship between the miRNA and its potential target sites. The orthogonality with respect to the transcriptome can be measured by the number of potential target sites a miRNA could bind to. The exact nature of a targeting relationship is unknown, as described in Chapter 2. It is known that a level of complementarity is required and the seed region requires stronger complementarity than the rest of the sequence.

This knowledge can be translated into a computational procedure as follows. Given a candidate miRNA sequence and a potential target site sequence, RNA hybridisation simulation can be performed and the result analysed to check for the level of complementarity at the seed and the rest of the sequence.

If this computation would be carried out between the miRNA sequence and a chosen set of subsequences of all of the transcripts, and one point would be scored every time a potential targeting pair was discovered, the sum of the points would give an *orthogonality score* for a miRNA sequence. In those terms the first orthogonality requirement means that miRNA sequences with lowest possible orthogonality score are sought after.

In the simplest case, an arbitrary but reasonable length of the subsequence could be chosen. This length would have to allow for bulges in the secondary structure of the pairing. The subsequences could be chosen in a sliding window procedure along all of the transcripts. The timescales involved in this approach are described further.

There are 140 millions of nucleotides in the human transcriptome (NCBI's cDNA database of confirmed and putative transcripts, version 36.54). Assuming that the sliding window will jump every five nucleotides and taking into account that the transcriptome is divided into transcripts, that will give about 25 millions of subsequences to check against. RNA hybridisation is computationally expensive, so optimistically assuming that analysing one miRNA:subsequence pair would take 1 millisecond on a modern CPU it would take about 7 hours to score one miRNA candidate.

The number of potential miRNA candidate sequences is 4^n, where n is the length of the miRNA. It is easy to see that the problem has exponential complex-

ity in the length of the miRNA. Even though it is a kind of static exponentiality, because the n is set, the numbers involved are large. For technical reasons mentioned in Chapter 6 the length of the miRNA has been set to 22, so there are $4^{22} = 1.76 \times 10^{13}$ candidates. This would amount to 1.4×10^{10} years.

The naive approach is clearly intractable, and the above solution does not address the second orthogonality requirement yet. Consequently, an approach is described to fulfilling all of the requirements using reasonable computational resources.

3.3 Design and implementation of software

The two orthogonality requirements are not independent of each other and sometimes work at cross purposes. Sequences that are orthogonal with respect to the transcriptome may tend to be similar to each other and therefore a simple pick of the top sequences fulfilling the first orthogonality requirement will result in an ensemble that violates the second orthogonality requirement.

The workflow presented in this chapter interleaves the steps designed to ensure both of the requirements. The top-level view of the process is presented in Figure 3.1.

All sequence computations were conducted in the DNA nucleotide space, so the T symbol (thymine) and not U (uridine) is used also when RNA is considered.

3.4 Choice of seeds

As mentioned in Section 2.5, mismatches in the seed region are the most conducive to making a potential miRNA:target site pair inoperable. Since the seed is the most constraining part of the miRNA sequence, a heuristic has been decided upon whereas the candidate seed sequences are chosen first, and the remainder of the sequences is added later. The aim is to find seeds that have poor complementarity with the transcriptome. The algorithm is fast enough to check all possible seeds, since there are only $4^7 = 16,384$ possible RNA 7-mers. However, not all of the possible seeds were taken to the scoring phase. The seeds with subsequences of

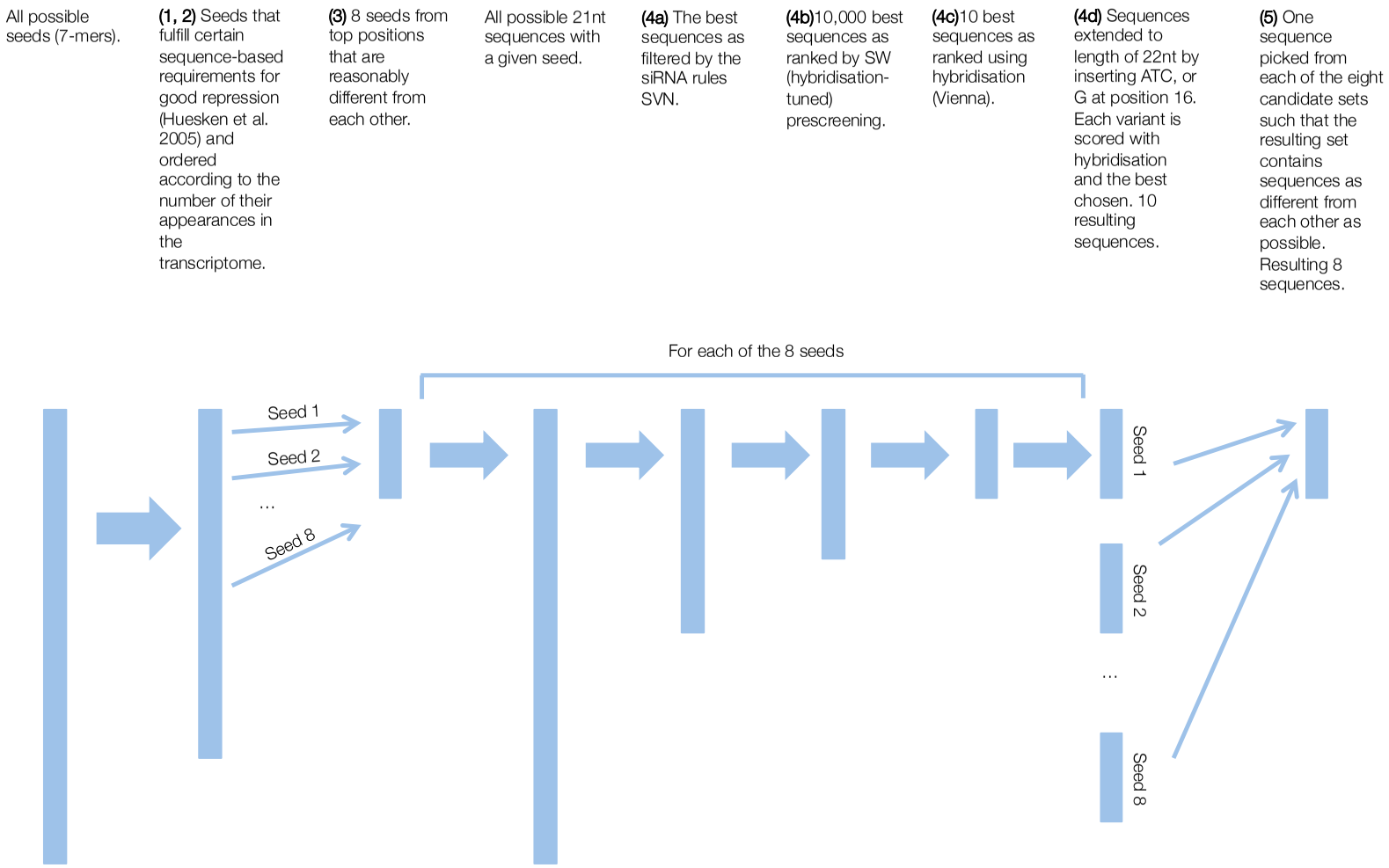

Figure 3.1: **The sequence of heuristics employed to select the synthetic miRNAs.** The steps are designed to employ constraints on the possible sequences in order to reduce the number of possible candidates. The numbers refer to the workflow points in the Introduction section.

four of the same consecutive nucleotides were excluded to avoid possible DNA secondary structures, leaving 15,552 candidates.

3.4.1 Seed scoring

The seeds were scored with a sliding window algorithm. A candidate was matched against each position in each transcript and the number of matches and mismatches between nucleotides on the same positions was calculated. The matching rules followed the Watson-Crick complementarity, so matching pairs were: A:T, C:G, and G:T. G:T was included because RNAs can also form G:U base pairs albeit not as stably.

Each position where there were at most two mismatches between the candidate seed and the transcript was considered a potential target site and the candidate would receive one penalty point. The seeds were then sorted in an ascending order by their accumulated penalty points with best ones on the top of the list.

Not surprisingly, the score is highly correlated with G/T content as those two nucleotides can be a part of two different pairs (Figure 3.2).

3.4.2 Applying the siRNA design rules to the seeds

The siRNA design rules apply to the seed as much as they do to the rest of the sequence. Therefore, it is important to try to apply them before the seeds are fixed. At this point, a simple rule based on the statistical analysis of Huesken et al. (2005) (Supplementary material) was applied.

The most statistically significant findings regarding the seed region were that adenine and thymine were favoured on positions 1, 2, 4, 5, and 6 of the seed. Applying those strictly would be too restrictive and unnecessary, so the filtering rule was that A or T was present at at least two positions mentioned above.

The remaining seeds had all scores below 4 million (under the dotted line in Figure 3.2), which removed the need for further restriction on the score.

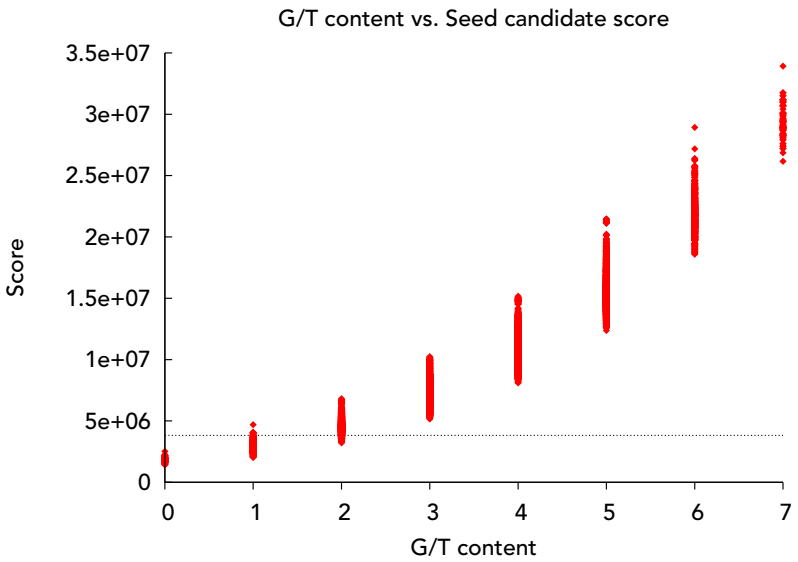

Figure 3.2: A seed's score (Section 3.4.1) is correlated with its G/T content with R = 0.949, which is related to the fact that both G and T can partake in two different RNA hybridisation pairs. Seeds compliant with the simple siRNA design rule fell below the dotted line.

Table 3.1: The Smith-Waterman scoring matrix that models RNA complementarity. This scheme is based on a simple model of RNA hybridisation where the Watson-Crick complementarity pairs—A-T/U and C-G—score high, and the wobbly G-T/U pairs are accepted but have weaker binding. Opening a gap was treated as a mismatch (-3), but continuing the gap was penalised less (-1).

	A	T	C	G
A	-3	2	-3	-3
T	2	-3	-3	0
C	-3	-3	-3	2
G	-3	0	2	-3

3.4.3 Choosing orthogonal seeds

The top seeds tend to be similar to each other, which violates the second orthogonality requirement. The choice of seeds from the candidate should maximise their pairwise distances.

Definition S is the set of all seed candidates.

Definition For any $s_1, s_2 \in S$, the distance between s_1 and s_2 is $SW(s_1, s_2)$, where SW is the complementarity measure between s_1 and s_2 based on the Smith-Waterman algorithm (Smith and Waterman, 1984). The Smith-Waterman scoring matrix was chosen to model RNA complementarity rather than similarity between two sequences in order to take into account the G:T base pairs (Table 3.1). In this set up the possible score values for 7-mers range between -21 (no complementarity) and 14 (full complementarity). In order to score correctly, the Smith-Waterman algorithm is performed on one of the two seeds and on the complement of the other.

Maximising the pairwise distances can be defined in several ways. For instance, the mean of all pairwise distances could be treated as the measure of how good a set of seeds is. However, the second orthogonality requirement states that any of the miRNAs should not interfere with each other and a mean could hide similar pairs in the ensemble. Thus, the approach taken here is that a set of seeds is as good as its worst pair, so subsets of the set of candidates are given a score which is equal to the distance between their most similar elements. Formally:

Problem For a given integer $k > 0$, find set $K \subseteq S, |K| = k$ such that

$$max\{SW(s_1, s_2) : s_1, s_2 \in K, s_1 \neq s_2\} \tag{3.1}$$

is minimal.

The above problem can be redefined in graph terms.

Definition Let $G = (V, E)$ be a clique graph, where V is a set of seed candidates, and E is a set of undirected weighted edges with weight d such that

$$E = \{(\{u, v\}, d) : u, v \in V, u \neq v \wedge d = SW(u, v)\} \tag{3.2}$$

The sought solution is a subclique of G.

The problem has a factorial time complexity as the number of all k-element subsets of a $|V|$-element set is $\binom{|V|}{k} = \frac{|V|!}{k!(|V|-k)!}$.

In order to deal with the complexity without sacrificing accuracy, G can be reduced by removal of edges with weights greater than a given cutoff. Accuracy is not lost, because any cliques containing edges that signify high level of complementarity would not have been of interest in the first place.

Such reduced graph can be subjected to the Bron-Kerbosh algorithm that finds all maximal cliques in a graph (Bron and Kerbosch, 1973). The value of the accepted complementarity cutoff was chosen empirically to be -1 by running the algorithm multiple times starting with low values (sparse graphs) and steadily increasing until cliques of size 8 appeared.

Out of all 47,577 resulting 8-element cliques the one with the lowest average complementarity score was chosen. Table 3.2 shows the chosen seeds and their place in the full miRNA sequences. Figure 3.3 shows their relative complementarity.

The following steps will instantiate the remaining nucleotides denoted by 'N' in Table 3.2 ensuring the two orthogonality requirements and compliance with the siRNA design rules. This leaves 15 nucleotides, thus $4^{15} = 1e9$ possibilities for each of the eight sequences.

As mentioned above, the complementarity scores were calculated using the Smith-Waterman algorithm with chosen scores that would sensibly reflect RNA

Table 3.2: The eight chosen seeds. The sequences are the miRNA guide strands.

Name	Sequence	Score
MT1	N**AAAGAAC**NNNNNNNNNNNNNN	3754837
MT2	N**AAATCCC**NNNNNNNNNNNNNN	2627539
MT3	N**ACCGTAA**NNNNNNNNNNNNNN	3744138
MT4	N**ACGCCTA**NNNNNNNNNNNNNN	3826865
MT5	N**ACTAACA**NNNNNNNNNNNNNN	2757761
MT6	N**CAACAAG**NNNNNNNNNNNNNN	3284993
MT7	N**CACCTCC**NNNNNNNNNNNNNN	3802587
MT8	N**ACCCGAT**NNNNNNNNNNNNNN	3343310

hybridisation. The Smith-Waterman algorithm produces a matching between two sequences that can be interpreted as hybridisation. In this scheme gaps are penalised and canonical base pairs (A:T/U, C:G) are favoured over wobbly G:T/U pairs. It would have also however been possible to use the free energy of simulated hybridisations between the seed candidates as a score. Such scoring would have been better grounded in the current state of the art understanding of hybridisation. Performing the same procedure, but with complementarity scoring based on free energy prediction estimated by the RNACofold utility of the Vienna RNA package (Hofacker, 2003), the eight chosen seeds would have been (with seed scores): AAATACC (2702627), AACATCC (2833154), ACAAATC (2689600), CAACGAC (2550589), CAATAAC (2702531), CATAAAC (2821019), AACTCAA (2780796), CACTAAC (2697110).

In spite of the suboptimal complementarity scoring method chosen, the experimental results did not show any violation of the inter-ensemble orthogonality.

3.5 Application of siRNA design rules

As mentioned in Section 2.6, efficiency of RNAi regulatory sequences is significantly affected by the sequence alone. The most established reason for that is the mechanism of the guide strand choice from the duplex by the RISC, which seems to depend on the relative stability of the two ends of the duplex. Other possibilities include stability of secondary structure of the guide strand and the G/C content.

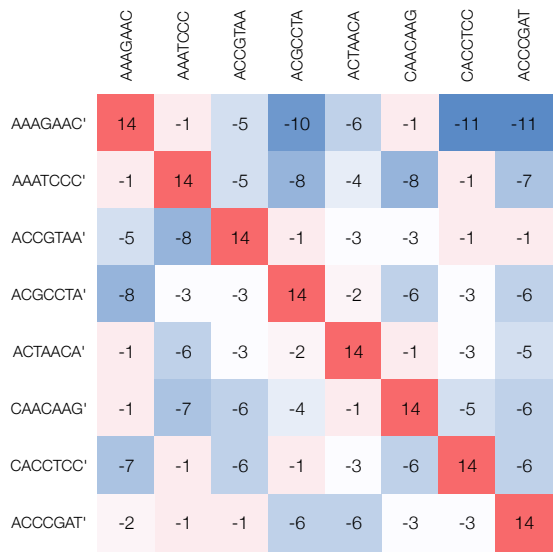

Figure 3.3: **The seed complementarity matrix.** This figure shows the relative complementarity between the chosen seeds (Section 3.4.3). The columns represent the seeds, while the rows represent their complementary sequences. High numbers (red) denote high levels of complementarity. Low numbers (blue) denote low levels of complementarity.

As the exact molecular mechanisms remain unknown, the most effective ways of prediction of siRNA/miRNA efficiency are the black box machine learning techniques. Enough data exists for the problem to be approached statistically.

In the next step of the procedure the candidates for entire miRNA sequences are filtered, so their efficiency in isolation is predicted to be high.

Huesken et al. (2005) successfully trained a neural network on a dataset of 2,431 siRNAs and their respective efficiencies. However, McQuisten and Peek (2009) found that Support Vector Machines (SVM) perform better than neural networks or linear regression in predicting the siRNA efficiency. Here, the same dataset of was used to train two SVMs—one more accurate but more complex and slower and another one that was simpler and faster with similar accuracy.

The first one was encoded following the Position Specific Base Composition (PSBC). For each position in the sequence four binary features were created, one for each possible nucleotide. A feature would equal 1 if the corresponding nucleotide existed at that position and 0 otherwise.

The second encoding took n-grams as well as PSBC into account. This encoding, implemented by McQuisten and Peek (2009), stems from the idea that efficiency of an siRNA can depend on sequence motifs rather than just isolated nucleotides. A feature is created for any possible sequence of length n and it is assigned value 1 if it is a substring of the siRNA sequence, and 0 otherwise. This encoding has not been made position specific in order to avoid the combinatorial explosion of the feature number. McQuisten and Peek (2009) found that PSBC together with n-grams for n-grams of lengths between 2 and 5 (inclusive) performed best, hence this combination was chosen here as well.

The PSBC+ng25 encoding generated a relatively large model and prediction was much more computationally expensive than with just PSBC.

The training and prediction of an epsilon regression SVM was done using the libSVM library, version 3.11(Chang and Lin, 2011).

First, using the C++ implementation of the command line tools the accuracy of training for both encodings was measured by using 10-fold cross validation as below.

```
$ svm-train -s 3 -v 10 PSBC.txt
$ svm-train -s 3 -v 10 PSBC+ng25.txt
```

The option `-s` selects the SVM type (here, epsilon-SVR), the option `-v` enables cross validation (here 10-fold). The PSBC encoding allowed for accuracy R = 0.641, and the PSBC+ng25 encoding for R = 0.691. Attempts to optimise for training parameters did not yield better results. The close results meant that the PSBC model could be used to prescreen the candidates before applying the PSBC+ng25 model.

In order to further validate this correlation, a permutation test was performed. In both cases (PSBC and PSBC+ng25) the original training set was partitioned into a 90% training set and 10% (243 samples) testing set. An SVM was trained using the training set and then the resulting model was used to predict values of the testing set. The correlation coefficiens in those particular cases were 0.636 (PSBC) and 0.663 (PSBC+ng25). In both cases a 1,000,000 sample permutation test was performed and in neither case was a sample found that exceeded the tested correlations, thus putting the p values under 1E-6.

The training itself was done using the same tool.

```
$ svm-train -s 3 PSBC.txt
$ svm-train -s 3 PSBC+ng25.txt
```

The output of the training were model files that could then be used for prediction. As shown before, the PSBC+ng25 model is slightly more accurate than the PSBC. It is however larger and takes longer to execute. An average time to execute the PSBC+ng25 model on one sequence is $5.18\,\text{ms} \pm 0.28\,\text{ms}$, while the PSBC is an order of magnitude faster at $0.279\,\text{ms} \pm 0.0169\,\text{ms}$ seconds per sequence (Measurements taken on a modern personal computer as of 2012).

For each chosen seed there are 4^{15} possible sequences to screen. The ones that have at least four consecutive nucleotides can be discarded quickly, but it does not markedly reduce the number of candidates.

The problem of screening the candidates was made tractable by prescreening them using the simpler and faster PSBC model and then reducing the list further with the full PSBC+ng25. Additionally, the Kaichu cluster computing framework (Chapter 4) was used to run the calculations in parallel on student lab machines. The resources available at their peak would be around 700 modern computing cores, but due to high intermittency of availability typically would not exceed 300.

Each cluster node was given 1,000 candidates to screen at a time. It would then run prediction using the *svn-predict* utility from libSVM first with the PSBC model. The candidates with predicted efficiency of 0.8 or greater were then screened using the PSBC+ng25 model and the ones with predicted efficiency greater than 0.9 would be returned to the master application that would then collate the results.

The number of the resulting candidates varied between 536,428 and 3,586,007 depending on the seed. For further calculations only highest-scoring 500,000 candidates for each seed were carried forward.

3.6 Scoring the full sequence miRNA candidates

Having chosen the seeds and filtering the full sequence candidates by applying the siRNA design rules the remaining sequences had to be assessed based on the number of hybridisations to the transcriptome that were likely to form working miRNA:target site pairs.

As stated in Section 2.5, such workable hybridisation would tolerate up to two mismatches in the seed region and up to five (but not more than three consecutive ones) outside of the seed region.

The sliding window algorithm described in 3.4.1 identifies likely positions in the transcriptome where a miRNA could hybridise to form an miRNA:target site pair. For each such position a candidate target site can be extracted from the transcript sequence and a hybridisation simulation can be performed between this sequence and an miRNA candidate.

RNA hybridisation algorithms such as those implemented in the Vienna RNA package (Hofacker, 2003) output the secondary structure in a bracket notation such as this

```
ACGAUCAGAGAUCAGAGCAUACGACAGCAG&ACGAAAAAAAGAGCAUACGACAGCAG
..((((...))))...((...((...((..&.............))...))...))..
```

where matching brackets indicate nucleotide pairings. Such output can be automatically analysed for conformance with the abovementioned rules.

The number of times such hybridisation conforms to those rules would provide a score of an miRNA candidate. The higher the score, the worse the candidate is.

State of the art hybridisation algorithms are, however, computationally expensive and running around three million hybridisations for each of the 500,000 candidates for each seed would be intractable even using a medium-sized computing cluster.

3.6.1 Smith-Waterman prescreening of candidates

The Vienna RNA package (Hofacker, 2003) provides a hybridisation utility—RNAcofold—that performs hybridisation between two RNA sequences. This cal-

culation is based on the dynamic minimum free energy algorithm originally developed by Zuker and Stiegler (1981). The algorithm takes into account detailed information on the binding strengths between RNA motifs and is computationally expensive. In order to compensate for that, an another heuristic employed in this workflow was to use the Smith-Waterman algorithm (Smith and Waterman, 1984) as a substitute to prescreen the candidates before the RNAcofold is used for the final scoring.

The Smith-Waterman algorithm can produce an alignment that can be interpreted as hybridisation where matches are nucleotide pairs and mismatches are bulges and has already been used in this way as described in 3.4.3. However in this case more effort has been made to reflect the work of the hybridisation algorithm more closely and the Smith-Waterman algorithm parameters were trained for that purpose.

Finding Smith-Waterman algorithm parameters that will allow it to closely resemble the results of a state of the art RNA hybridisation algorithm is an optimisation problem. One of the most powerful optimisation techniques, if it can be employed, is the graduated optimisation (Rosenfeld et al., 1984).

Graduated optimisation is a technique initially developed for image analysis, but then found its uses in many areas of engineering (Afanasjev et al., 1989; Gashler et al., 2011; Ye et al., 2003). It aims to avoid settling in local minima by creating stages of progressively blurred fitness landscape. Ideally, the landscape in the most blurred stage will look like a basin where there is only one minimum that can be reached easily by using hill climbing. Each of the less blurred stages allow small refinements and by the time the original, sharp, landscape is reached the found location will be in the global minimum.

The original training set was created by randomly generating 1,000 21 nt long RNA sequences to stand as miRNAs and 256,000 38 nt long RNA sequences to stand as potential target sites. The hybridisation scores were calculated for each of the 1000 miRNA sequences against the 256,000 potential target sites using the Vienna RNA package's *RNAcofold* utility and by analysing its output against the targeting rules. Thus, a mapping was created from the miRNA sequences to their integer scores.

The fitness function would be the correlation between the scores calculated

with *RNAcofold* and those calculated by the Smith-Waterman algorithm parametrised with a given parameter vector.

The blurred stages were generated by calculating the same mapping but by halving the number of the potential target sites. So, the progressively blurred stages were created for sets of 128,000, 64,000, and down to 1,000 potential target sites.

In order to increase the chances of finding a good minimum a latin hypercube sample of 1,024 parameter vectors was generated using the *lhsdesign* function in Matlab. The range of initial values for each parameter was [-5; 5], with only integer values taken into account, but the parameters were allowed to leave that range during training. A parameter vector had twelve values—ten for each possible nucleotide pair plus the start gap penalty and the extend gap penalty.

The procedure was performed in 9 phases, one for each fitness landscape generated. The first phase involved the set of 1,000 potential target sites and all 1,024 parameter vectors. In each phase each of the parameter vectors would perform hill climbing. The neighbours considered in each step of hill climbing would all be parameter vectors that differ by one of the parameters with the difference equal to 1. That made for 24 neighbours at each stage. When no more improvements could be made for any of the parameter vectors, the next phase would occur. In the next phase, double the size potential target site set would be used, while the parameter vector set would be halved by eliminating the worst-scoring vectors. Table 3.3 shows the progression of the phases from the smallest set of the potential target sites (most blurred fitness landscape) to the largest (sharp fitness landscape) and from 1,024 considered parameter vectors to just 4 in the last step.

The best resulting parameter vector reached Pearson correlation of 0.65 and is shown in Table 3.4. Those parameters have been validated against a different random set of 256,000 target sites and a different set of 100 random miRNAs. The correlation between the scores achieved through hybridisation and with the Smith-Waterman algorithm was 0.669 at p = 2.4899E-14. A permutation test with 1,000,000 samples was performed where a correlation between one group was calculated against a randomly permutated second group. The mean value of the sampled correlations was very close to zero (1.8075E-04) and neither of the

Table 3.3: Progression of the graduated optimisation phases in training the Smith-Waterman parameters to mimic the Vienna RNA hybridisation algorithm.

Phase	**Target site set size**	**Parameter vector set size**
1	1,000	1,024
2	2,000	512
3	4,000	256
4	8,000	128
5	16,000	64
6	32,000	32
7	64,000	16
8	128,000	8
9	256,000	4

Table 3.4: The Smith-Waterman scoring matrix that models the behaviour of the *RNAcofold* algorithm of the Vienna RNA package. Open gap penalty = -5, continue gap penalty = 0. The correlation between the targeting scores achieved based on the true RNA hybridisation algorithm and and based on the Smith-Waterman algorithm with these parameters is 0.65.

	A	T	C	G
A	-2	1	0	-10
T	1	4	0	-10
C	0	0	1	3
G	-10	-10	3	3

samples exceeded 0.669 which estimates the p value at less than 1E-06.

Using the found Smith-Waterman parameters, all of the 500,000 miRNA candidates were prescreened to select the 1,000 best ones for each of the chosen seeds. Again Kaichu was used and, to speed up the calculations further, a global dynamic cutoff value was maintained that all of the computing nodes had access to and updated on periodically. The cutoff was the score of the current 1,000th best candidate. Thanks to this, a miRNA scoring calculation could be finished early once the score had exceeded the cutoff. The exact scores of the candidates beyond the best 1,000 were unimportant.

3.6.2 Hybridisation scoring of candidates

With 1,000 miRNAs left for each of the chosen seeds it was tractable to score them based on RNA hybridisation simulation provided by the Vienna RNA's *RNAcofold* utility.

In order to eliminate the overhead of the process start time, *RNAcofold* was used in interactive mode where it reads from the standard input and writes the results to the standard output. Those results were then captured by a Java wrapper, which analysed the bracket notation for compliance with the targeting rules and, if necessary, updated the score of the candidate.

Similarly to the prescreening calculation, a global dynamic cutoff value was employed to terminate calculations of particularly bad candidates early, however this time only the 20 best candidates were chosen for each seed.

3.7 Sequence extension

The dataset released by Huesken et al. (2005) was generated for siRNAs of length 21, and the siRNA design rules that stem from it are meaningful for sequences of that length. 21 was therefore a natural choice for the designed miRNA sequence length and the first pass of the miRNA discovery workflow generated sequences of that length. However the miRNAs used in the experimental setup described in Chapter 6 failed to work.

It was hypothesised that this was caused by the fact that the DNA construct used was originally meant for 22 nt miRNA sequences and shortened pri-miRNAs failed to process correctly. Indeed, Zeng and Cullen (2003) found that a bulge close to the cleavage site significantly interfered with the processing (most likely with the cleavage by Drosha) and very few miRNAs would be produced. The pri-miRNAs of the testing construct indeed have a bulge in that region and a shorter stem loop brought it even closer. To further test this hypothesis, one of the failing miRNAs was arbitrarily extended by one nucleotide, tested and had been found to work.

A more principled sequence extension process was designed to adapt the sequences to the test construct. Ideally the additional nucleotide should be inserted

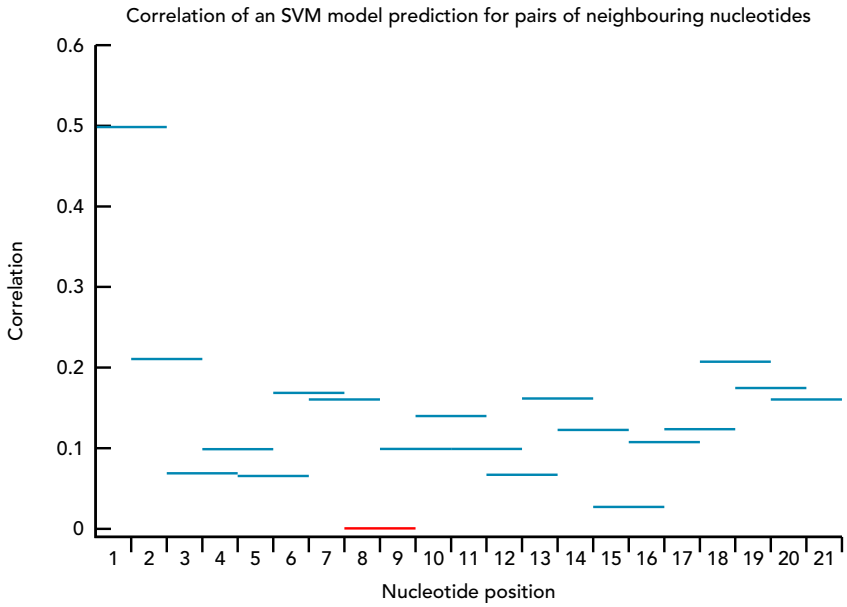

Figure 3.4: The pair of the 8th and 9th nucleotides of an siRNA contributes the least in terms of efficiency on the siRNA. Thus, it is likely that sequence extension by insertion of one nucleotide between those two should disturb the efficiency the least.

in the region that contributes the least to the efficiency of the regulator in light of the siRNA design rules. The Huesken et al. (2005) dataset was used again to calculate the correlation between substrings of length 2 at all positions and the measured efficiency. The results are shown in Figure 3.4. Based on that data the least sensitive insertion position should be between nucleotides 8 and 9.

There are four possibilities (A, C, T, and G) and this choice was made on the basis of the orthogonality score that was calculated for all possible elongated sequences.

3.8 Choice of eight candidates

At this point 20 candidates of similar quality for each of the chosen seeds remained, and the second orthogonality requirement could be applied once again, but to the full sequences this time. This problem was similar to the one described in Subsection 3.4.3 regarding seeds, but instead of drawing eight elements from

one set, the requirement was to draw exactly one element from each of eight sets. Again the goodness of a set of candidates was considered to be based on the similarity of the most similar pair.

More formally, given sets of sequences $MT_1, MT_2, ...MT_n$, a cartesian product of those sets

$$K = MT_1 \times MT_2 \times ... \times MT_n \tag{3.3}$$

a function

$$d : \bigcup_{k=1}^{n} MT_k \times \bigcup_{k=1}^{n} MT_k \to \mathbb{R} \tag{3.4}$$

which maps a pair of sequences to their Smith-Waterman difference, and a function

$$e : K \to \mathbb{R} \tag{3.5}$$

$$(mt_1, mt_2, ..., mt_n) \mapsto min\{d(mt_a, mt_b) : \forall_{a,b \in [1,...,n], a \neq b}\} \tag{3.6}$$

choose an element of K, such that e is maximal.

As a solution a backtracking algorithm was employed where sets of candidates were formed by adding candidates one by one from each set. Every time a candidate was added, its similarity to all the other candidates already in the set was calculated. The similarity was calculated in the same way as in 3.4.3. If any of the newly calculated similarities was greater than the score of the best set so far, the algorithm would backtrack and save time.

The best eight sequences chosen together with their orthogonality scores are shown in Table 3.6.

3.9 Off-targeting sequences

It is possible that sequences equal in terms of conformance to the siRNA rules will be more or less efficient depending on the number of their off targets in the transcriptome, however there is a question whether targets with likely smaller affinity to the miRNA that the intended target has could markedly divert the miRNA towards themselves. To test whether sequences with equally good siRNA design rules would differ in their efficiency depending on the number of their

Table 3.5: The eight chosen miRNA candidates with their orthogonality scores. The sequences are the miRNA guide strands. The seeds are in boldface.

Name	Sequence	Score
MT1	T**AAAGAAC**CAACACCCTACCTA	5,936
MT2	T**AAATCCC**CTATAAACAACCAA	3,489
MT3	T**ACCGTAA**ACAAATATACCCTA	7,766
MT4	T**ACGCCTA**ATAATAAACACCTT	8,681
MT5	T**ACTAACA**ACAACATCAACCAA	8,227
MT6	T**CAACAAG**CCATCCACAAACAT	9,664
MT7	T**CACCTCC**CTAAACAAATCGTA	7,878
MT8	T**ACCCGAT**AAAACACAAACCTA	4,876

predicted off-targets ideally a large number of sequences with different off-target scores could be produced and tested in the lab. However, the financial and time constraints as well as the fact that this work's main goal was to provide performant miRNAs prevented such large scale test that could provide statistically significant results.

Still, a preliminary test with two produced highly off-targeting sequences was devised with a prediction that they would perform badly. The starting point were two different seeds chosen from the back of the list described in 3.4.1. Those seeds—TTTCTTT and AGGTGGG—have a lot of matches discovered by the sliding window algorithm (2.8933073E7 and 2.2770162E7 respectively). Then, the procedure followed similarly with filtering the candidates that conform well to the siRNA design rules. Prescreening with the Smith-Waterman algorithm and scoring with the *RNAcofold* was not tractable as the cutoff optimisation could not be employed since it were the large off target scores that were sought after. Therefore, only 30 of the 500,000 candidates for each seed were scored using the *RNAcofold* as scores two orders of magnitude higher than those achieved by the MTX sequences were easily found, and the ones with the highest predicted off targeting were chosen. Similarly, those were extended by one nucleotide and from the four possibilities of extension the ones with the highest predicted off targeting were chosen. There was no consideration necessary for the second orthogonality requirement in this case. The two chosen off-targeting sequences are shown in Table 3.6.

Table 3.6: The two negative control miRNAs with their orthogonality scores. The sequences are the miRNA guide strands. The seeds are in boldface.

Name	Sequence	Score
OT1	A**TTTCTTT**GGTTTGTCACTGTG	498,685
OT2	T**AGGTGGG**GCAAGTGTGTTCGT	532,383

3.10 Summary

In this chapter the whole workflow for discovery of orthogonal miRNA sequences is presented on an example of the human transcriptome and eight found sequences. The two orthogonality requirements are stated. The first orthogonality requirement regards orthogonality with respect to the transcriptome and the second orthogonality requirement regards orthogonality between the ensemble of produced miRNAs.

At the end there is a brief description of how the negative controls were produced—miRNAs that have high off target scores and therefore are poor in terms of the first orthogonality requirement. This workflow follows a slightly modified procedure presented for seeking for good sequences.

Chapter 4

Kaichu Cluster Computing Framework

4.1 Introduction

Abundance of spare computational power from resources such as teaching machines at universities motivates researchers to develop their software to run in a parallelised fashion. Such resources can be utilised outside of the working hours, so the machines are fully available to the computations, but that still leaves a large potential untapped. With the advent of computational cloud services such as Amazon Elastic Compute Cloud 2 (Amazon EC2) cheap computational power is available to researchers. However, access to cheap or free computational power if often not taken advantage of. The most probable reason is that developing distributed systems that are correct and efficient, especially in an uncertain environment, is difficult, and the available solutions are unwieldy for small projects or fast-changing code bases.

This chapter introduces a cluster computing framework—Kaichu—developed to aid the parallelisation of the computations required in the main work in this project, but designed for general use. Kaichu was intended to be robust to uncertain environments such as pools of computers that often become available and unavailable again and has a number of unique features, such as automatic job aggregation. The two main design goals were to enable rapid development of

parallelised software and efficiency.

In the following sections I review several available software frameworks, then introduce the principles on which Kaichu has been built in more detail; I speak about the technical design; introduce the main notions and features and finish with benchmarks.

4.2 Overview of cluster computing systems

In this section, three popular cluster computing frameworks are introduced as a context for Kaichu. From the original and still widely used Condor, to more modern batch Hadoop, to the most forward-thinking commercial GridGain. Kaichu draws inspiration from all of those and makes its place as a modern, open source alternative especially suited for computationally-intensive ad hoc applications running in highly uncertain environments such as office or student lab machines.

4.2.1 Condor

Condor (Litzkow et al., 1988) is a batch computing system designed with cycle scavenging in mind. The idea of cycle scavenging is based on a fact that in most cases users of office computers do not utilise all of their computational power. A background, unobtrusive task could use the spare power to perform jobs retrieved from the network and thus researchers and companies could increase their computational ability at almost no cost.

Thus, Condor is a complex system of daemons that are always present on the worker nodes and a master node that coordinates them. Any node can be a submit node which means that it can submit jobs, which will be retrieved by the worker nodes for execution.

Technically setting up a computation with Condor entails design and implementation of worker applications, and setting their distribution and data division using the Condor run scripts and input files. After the system performs the computations, the results are returned to the submit node in the form of output files. Those can be then aggregated.

Unfortunately, Condor does not make provisions for running a computational

master node that could automatically detect and combine the result data. Therefore it is difficult to implement applications that require periodic aggregation of the results. For instance, a genetic algorithm is run in rounds whereas the fitness function computation can be parallelised, but once every generation a global decision has to be made which members of the population are to be discarded and which mutated/mated and put through to the next generation. The engineering effort required to implement such application with Condor makes the process unwieldy.

Condor does not support automatic aggregation of jobs and the task of tuning the size of the jobs falls on the engineer. This is made worse by the fact that the start time in Condor can be in minutes which forces relatively large jobs. On the other hand however, large jobs are not necessarily efficient as a node failure means that the whole job has to be repeated on another node. Condor does support chekpointing which means that a failed job could start from a certain point of the execution, but checkpointing requires the worker applications to be compiled using a special compiler which further complicates the engineering effort.

The fact that Condor requires development of separate worker applications poses two engineering challenges. One is that an existing single-machine application has to be carefully parcelled into the worker and the master part. Such division of the code base can be a significant undertaking and leads to the second problem which is that it may be unclear where point of division should lay and, given that the computational requirements can change, it can be a labour-intensive process to keep the application factored optimally.

Furthermore, each job requires either the network transfer of all code necessary to compute it, or some of the software has to be preinstalled on the worker nodes. The former further exacerbates the inefficiency, the latter complicates the deployment.

The above arguments show that Condor provided significant advancements at the time of its conception (which resulted in its popularity), but it can be unwieldy in use and inefficient.

4.2.2 Hadoop

Hadoop (Shvachko et al., 2010) is an open source computing platform modelled after Google's proprietary MapReduce (Dean and Ghemawat, 2008). Hadoop is a batch system designed to deal with large data sets and provide high degree of fault tolerance. A part underlying the computation is the Hadoop Distributed File System (HDFS), which is responsible for automatic data replication across the nodes to account for node failure. Hadoop makes an effort to localise computations together with the data.

Hadoop focuses on implementation of the Map Reduce paradigm whereas data for any computation is divided into chunks, each chunk can be independently computed on different machines (map), and then the results are aggregated (reduced) into a final result. In Hadoop also the reduction step is parallelised.

MapReduce is especially well suited for querying large, static, unstructured databases, PageRank (Page et al., 1999) being the prime example. As such, it is suited for large clusters of commodity machines, but prefers dedicated deployment and is not ideal for ad-hoc applications.

4.2.3 GridGain

GridGain is a commercial compute grid and data grid system geared towards complex implementations of distributed computational systems. GridGain, unlike Condor or Hadoop, is an in-memory, near realtime system and, out of the three, is the closest in that respect to Kaichu. Unfortunately, at the time of writing, the free Community Edition is not available any more.

GridGain is JVM-based and supports Java, Scala, and Groovy. The computing node workers have to be deployed manually, but they use IP-multicast discovery to discover each other automatically. GridGain implements what it calls a zero-deployment—any task code that has to run on the workers is loaded automatically and at runtime, so that the deployment of GridGain is identical independently of application.

In terms of features important for efficiency, GridGain implements features such as several methods of load balancing (round-robin, random, and adaptive), redundant mapping whereas one job can be executed by more than one node to

ensure liveliness of the application in case of node failure. Checkpointing allows for saving of partially complete computations, so in case of a node failure a job can be restarted on another node from using the checkpointed state, and not all over again.

GridGain is suited to perform with a number of data storage solutions and ensures efficiency by data and computation collocation—the computations are done as close as possible to the source of the data in terms of network topology to decrease the effects of network latency.

4.3 Approach and requirements

4.3.1 Overhead of distributed computing

One of the reasons why distributed computation is difficult are the communication overheads between the nodes. Excluding expensive purpose built computing clusters, typically the machines are connected in an Ethernet network with speeds of 100 Mb/s or 1 Gb/s which are orders of magnitude slower than modern bus communications. Data intensive computations suffer from this speed bottleneck, but also from network latency. If the jobs sent to the nodes are too small, the overhead of network latency starts being significant and reduces the gains from distributing the computation. Additionally, batch systems such as Condor significantly add to this latency. Starting a job on Condor not unusually takes a few minutes or more. This problem is solved by tuning and aggregation of jobs. There are problems with this approach however. Jobs that are too large can be lost in an uncertain environment. For instance, a student may restart the computer the job is running on. The system might also dynamically tune the jobs it is sending for performance and having large jobs will mean worse responsiveness. Furthermore, manual tuning of the job size can be time consuming for the developer and in rapidly changing requirements (such as is often with research applications) this effectively limits what can be done.

Kaichu aims to solve this problem from both sides. It is developed to minimise the overhead of the system, so that small jobs ($\sim$ 1 sec) can be run with very little effective overhead. On the other hand Kaichu provides an automatic aggregation

feature that lets it adapt the job size automatically at runtime.

With regards to the data intensive applications Kaichu offers a notion of a resource - any data that is static across all jobs that is configured in code and is automatically sent to the worker nodes as they need it, but only once per computation. It is done using a peer-to-peer network between the workers, so even large amounts of data can be distributed quickly.

4.3.2 Rapid development

In a solution with a rapidly changing requirements it is crucial that the cluster framework supports the developer and allows for rapid changes. In Kaichu there is no need for external configuration files or interfacing with files, etc. Everything happens in the Java code space and Kaichu provides a solid feeling of running the tasks on a large supercomputer.

Another aspect of that is fault tolerance. Computer networks are unreliable, the nodes are often not under full control of the user and also could fail. For the master node it is not always obvious what is happening with the jobs—whether they are still running or should be restarted. All this complexity makes ad hoc distributed computing solutions prohibitive in all but the biggest projects. Kaichu deals with faults automatically and has provisions to ensure liveliness of the system. As long as any worker nodes are running, the whole computation is guaranteed to complete.

4.4 Technical design

4.4.1 Java Remote Method Invocation

Java Remote Method Invocation (Java RMI) Oracle (n.d.) is an object-based remote procedure call API available in the core Java. It supports object serialisation and deserialisation for network transport and allows transfer of code as well as data between the distributed nodes and performs distributed garbage collection.

Remote objects are made accessible through the network by RMI resource registries. A registry must be run by one of the machines in the distributed

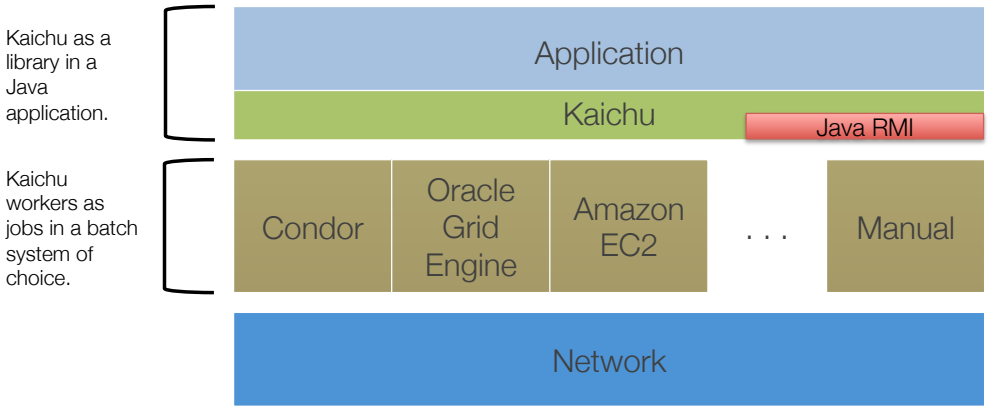

Figure 4.1: **The Kaichu stack.** Kaichu is imported into the application, provides the APIs and the runtime to manage the distributed computation. Kaichu itself is based on Java RMI for communication with the worker nodes. The worker applications have to be deployed either manually or using any of the batch systems. Once the workers are running, Kaichu takes over the communication and job distribution is performed efficiently.

system. All of the nodes have to know where it is running and be able to access it. The remote objects are registered in the registry to be accessed.

In order to enable code transmission between the nodes, an HTTP server must be run by the providing machine that gives access to the code base.

4.4.2 Overview of the network architecture

The master application imports the Kaichu library and uses its API to create task, jobs, etc. The framework provides the worker applications. Bootstrapping the system requires only starting the worker applications on the worker machines and providing them the hostname of the master application as a parameter.

Starting the worker applications manually can be tedious, so one of the batch cluster systems such as Condor or Oracle Grid Engine can be utilised to do it. In such configuration, the workers are treated as jobs by the batch systems, but they run as long as Kaichu needs them to perform Kaichu jobs. Any additional utilities or libraries that are needed to perform the computations can be transferred using the run scripts of the batch systems.

Kaichu can also run on computational cloud services such as Amazon EC2. A

virtual machine instance can be created that, upon startup, retrieve the master hostname from a web service and automatically start the Kaichu worker for each available CPU core, feeding it the hostname as parameter. dupa

4.4.3 Progress of the computation

The workers do not have any knowledge of the jobs they will be computing at the beginning as they are not aware of any application-specific code. Upon startup, a worker attempts to connect to the master to register themselves there. If the attempt fails, it will try again after a random interval of between 25 and 30 seconds.

When the user code in the master application initialises Kaichu, it starts an RMI process, registers its main remote object there to be accessed by the workers, and starts a simple HTTP server thread to enable GET access to the code. Upon receiving the registration message from the worker, the master adds it to the list of idle workers.

In the simplest case, the user code then should register one or more tasks (more in Section 4.5.1) which define the code that the worker must perform to compute a job in the Kaichu Task Registry. Then the user code can submit jobs to the Dispatch Queue (more in Section 4.5.3).

If there are jobs pending, Kaichu will connect idle workers in a round robin fashion and send them jobs. A job will identify the task that it can be computed with. When a worker receives a job, it checks whether it has the task definition in its local Task Registry. If it does not, which always happens at the beginning, it will make a separate call to the master to retrieve the task. After receiving it, the worker will start running the job immediately and return the job to the master. Next time the same kind of job is sent to the worker, it will already know how to compute it.

Thanks to this mechanism afforded to Kaichu by Java RMI, there is no need to preset the workers.

4.4.4 Job Starter threads

In the initial design Kaichu would start a separate thread that acted as a proxy and a monitor for any active worker node. That was however not scalable as the number of thread that a process can start is limited and on Ubuntu Linux the limit appeared already when using an order of 600 nodes. The first version also kept an open connection to any node that was in the process of performing a job.

Currently there is a tunable, but small number of starter threads (default at 10) that poll the jobs submitted to Kaichu, send them to the worker nodes and the RMI call ends. When a worker completes a job, it makes a call to the master node to send the results and goes idle until it is sent more work.

4.4.5 The Poke thread

It is important to monitor what the workers are doing in case the network connectivity to them is lost or the job fails to complete for any reason. The Poke thread is a monitor thread that calls the workers in a round robin fashion. If a worker is unresponsive, the master node declares it inactive and the job that had been sent to it is put on the front of the dispatch queue for rescheduling. It is ensured that if that same worker regains connectivity and sends the result, it is discarded and does not confuse the client application.

4.5 Features

This section presents a brief rundown of features and other user-facing elements that Kaichu API offers and how to use them.

4.5.1 Tasks

A task is a class that implements the `KaichuTask` interface to state the code that is run as part of a job. The interface is simple and takes one argument—the input data for the task, an object of a class implementing the `KaichuParameters` interface—and returns an object of a class that implements the `KaichuResponse` interface.

In order to be used, a task has to be registered in the Kaichu Task Registry. More than one task can be used in an execution.

4.5.2 Kaichu Manager

A `KaichuManager` object represents Kaichu to the user. It is used to initialise Kaichu and set parameters such as the debug output or the number of starter threads. The two methods—`start()` and `stop()`—allow the user to halt and resume Kaichu. In order to shutdown and invalidate the manager object, the user should call the `terminate()` method. Kaichu Manager can be also used to retrieve the Dispatch Queue and access to other features described below.

4.5.3 The Dispatch Queue and Jobs

A `KaichuManager` object holds an instance of the `DispatchQueue` class. This class allows for submission of jobs for execution as well as clearing any scheduled jobs. The `submitJob()` method takes one argument of `KaichuJob` type. A `KaichuJob` is created given an object of class implementing the `KaichuParameters` interface and the class object of the task class that is used to compute the job. Using the class object instead of, for instance, a string to tag a job with a particular task is less prone to programmer errors.

4.5.4 The Job Creator

In some applications there is a very large number of jobs to be computed that could not fit in the memory of the master application machine. In such cases, instead of trying to submit all jobs to the Dispatch Queue, they can be created on the fly, as requested by Kaichu.

The `KaichuJobCreator` interface is provides for that regime. A class that implements this interface has to override the `createNewJob()` method that has the `KaichuJob` return type. If there are no more jobs to submit at the moment, this method should return `null`. The Job Creator has to be then registered with the Kaichu Manager as the current job creator.

The `DispatchQueue` is in fact a convenience class that implements the `JobCreator` interface and maintains a queue of jobs. It is the default job creator.

4.5.5 Response retrieval

`KaichuManager` implements four methods for retrieval of responses. They are all variants of the `getNextResponse()` method. The plain version returns the responses in the order in which they had been received by the master application and blocks the execution of the thread it is called from if there are no more responses at the moment. The other three variants allow all combinations of blocking/non-blocking behaviour and retrieving the responses in the order in which their dual jobs have been submitted or in which the responses have been received.

The user code has to know how many jobs in total were going to be submitted, so it does not wait for more than that. Because Kaichu uses the pull pattern with the Job Creator, it has no way of telling when all jobs have been submitted and therefore it does not know when the computation is finished.

4.5.6 Automatic job aggregation

As mentioned earlier, Kaichu provides for automatic job aggregation. During the execution, the average job execution time is monitored by calculating a running mean with window size of 20. If the execution time of an aggregate is found to be outside of the range between 2 and 6 seconds, the aggregate size is adjusted to try and keep the execution time in the middle of this range. A new adjustment is not taken until there are at least 20 aggregates of the new size returned.

Automatic job aggregation is especially useful in applications where the job sizes steadily change as the computation progresses.

The automated aggregation policy gives very good results a typical case, but in case the user wants to turn it off and provide their own aggregate sizes, it is possible through APIs on the `KaichuManager`.

4.5.7 Rounds

Some algorithms require multiple turns of map and reduce. For instance, in a genetic algorithm, calculation of the fitness function can be parallelised—but between every generation when the best members of the population have to be selected—all of the information has to be collated in one place and global decision has to be made. Kaichu provides for such algorithms with the notion of rounds.

4.5.8 Resources and peer-to-peer distribution

It is sometimes necessary that the jobs require an access to a larger, static data. It would be prohibitive to send all of such data with each job. It would be possible to have it referred from a task which then would cause its transfer to the workers, but that is not the most elegant solution and would not allow efficient mixing different tasks and resources.

Examples of such resources in the miRNA design project include the transcriptome, or precalculated positions at which a seed matches the transcriptome.

In Kaichu, any object can be registered in the Resource Registry. In a Task's `execution()` method this resource can be referred to by its identifier string. The distribution of resources is done in the same lazy evaluation fashion as is the distribution of tasks—they are downloaded to workers only when explicitly needed.

Large number of worker nodes can very easily put a strain on the master's link to the network. If the master was to sent a 100MB file to 300 workers, it would have to send around 30GB of data and that can take a nontrivial amount of time.

To compensate for that, a feature for peer-to-peer distribution of resources has been implemented. In this schema, the master acts as a seed for the resource data. It sends any resource to at most two workers at a time. At the same time, any worker that successfully downloads a resource, communicates it to the master. The master keeps a list of such workers and redirects new requests to as many worker nodes as possible. As a result, the resources can be distributed at speeds that grow exponentially.

4.6 Benchmark

A benchmark has been conducted to assess the overhead that Kaichu may introduce.

4.6.1 Set up

The set up involved dual core, Intel i5 processor student lab machines in the School of Information Systems, Computing and Mathematics in located in five labs across four buildings across the Brunel University campus. The LAN was rated at 100Mb/s. The worker machines run Windows 7, while the master run a Ubuntu Desktop Linux distribution. All of the machines were a part of a Condor cluster and it was used to start the workers. Each machine could run at most two workers—one per each core.

4.6.2 Procedure

Each job consisted of solely a 30 seconds wait, but a randomly generated data of sizes 5KB, 50KB, 500KB, or 5MB were tagged along with each job. Another parameter that was varied was the number of workers with 1, 10, 50, 100, 300, and 600 workers at a time.

For each data point, the number of jobs generated was $10 \times$ #workers. The theoretical ideal time for each run would be $10 \times 30\,\mathrm{s} = 300\,\mathrm{s}$. During the execution, the wall time was measured and at the end the ratio of ideal time to actual wall time was calculated. With no overhead, this value would be equal 1 each time.

The measurements were done in three repeats.

4.6.3 Results

The results are shown in Figure 4.2. The lines for the two smallest job data sizes indicate no overhead all the way up to 600 workers used. The case of 500KB job size shows a small drop off, but not below 0.9.

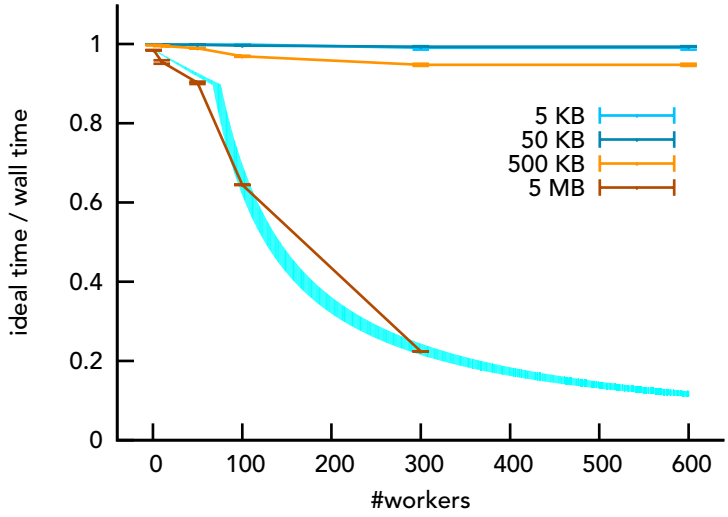

Figure 4.2: Ratio of ideal run time to the actual run time. Three runs were made for each data point and mean and standard deviation are shown. Kaichu's overhead is negligible in this set up.

A large drop off was observed in the case of 5MB jobs. This was however predictable. The light blue shade represents a theoretical curve for the 5MB case assuming real network speeds of between 10 and 11MB/s. The phase change in the curve marks the situation when there are so many workers that the master spends all of its time sending the jobs.

The last data point in the 5MB case was not calculated due to the difficulty in holding an exact number of 600 workers for extended periods of time on the test cluster.

4.7 Kaichu Monitor

When then number of workers in a cluster increases, any console output proves inadequate to monitor the overall state of the computation.

The Kaichu Monitor application was written to provide an at-a-glance view of the computation status. The application connects to the master application and polls for updates on all kinds of data. It shows summary statistics such as the number of jobs generated, number of jobs done, and the total number of active

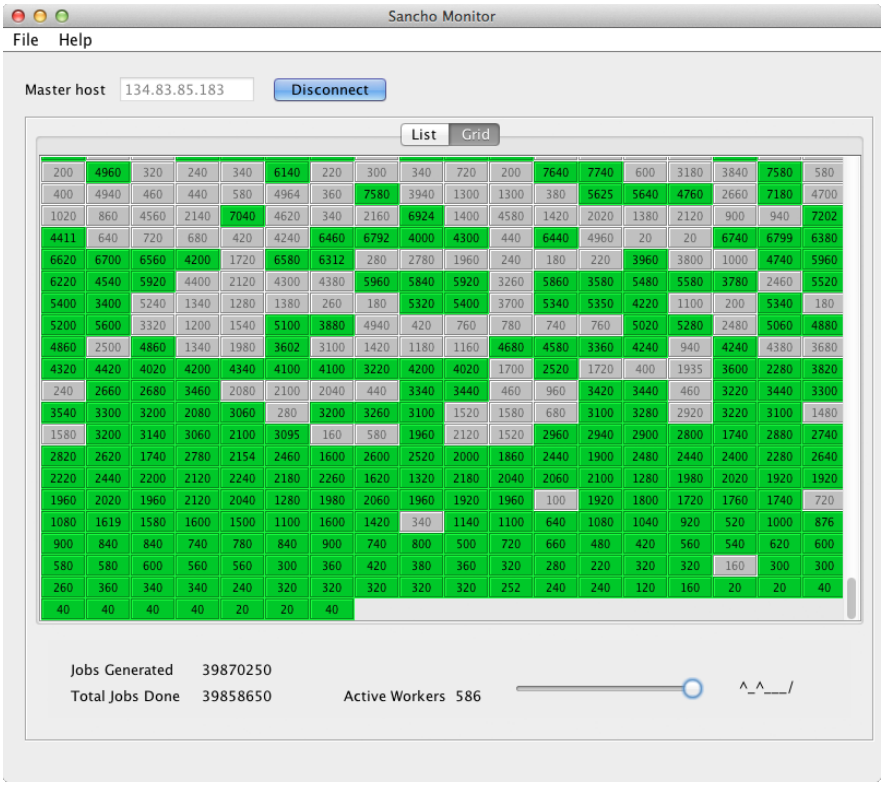

Figure 4.3: The Kaichu Monitor application provides a quick insight into the state of the cluster and each worker. The grid view shown in this figure represents each worker as a coloured rectangle. Both the colour and label of the rectange provide information about the state of the worker.

workers.

Additionally, the application provides a list view and a grid view where the state of each worker can be seen. The grid view (Figure 4.3) shows each worker as a rectangle. The label and colour of the rectangle indicate the state of the worker. For instance, a blue colour indicates that the worker is downloading a resource. At the same time letter 'W' indicates that the resource is being downloaded from another worker, while letter 'M' indicates that it is being downloaded from the master. Most of the time, the label displays the number of jobs that the worker has completed. With this label, light red colour means that the worker is idling, green—that the worker is computing, and grey indicates an inactive worker.

The list view provides all of the information shown in the grid view with additional field to show the last time it has been touched with the Poke thread.

4.8 Conclusion

Kaichu was built with the intention to provide a full-featured and yet simple to use master-worker cluster computing system. It was meant to effectively shield the user from the difficulties of distributed computing and allow for changes in code to be convenient. In the experience of this project Kaichu fulfilled those requirements. However, building a novel cluster computing system was not the focus of this project and therefore its robustness or ease of use were not formally tested.

Apart from the above-mentioned qualifications, there are several features that would further improve Kaichu's utility.

In an environment where the nodes are likely to fail, the liveliness of the system suffers if such failures occur at the end of each round. Failures at that time mean that the system has to detect the failure, resend the job to another worker, and wait for its completion again. This can make for noticeable time loss when the job size is relatively large comparing to the round execution time. A potential solution to this problem would be to overdrive the system towards the end of a round by sending the last job or several last jobs in the round to more than one worker. The probability that all of the two or more workers would fail would be much smaller. In order to implement that feature, several provisions would have to be made. For instance, Kaichu would have to be aware which of the jobs are the last ones in the round and be able to deal with multiple responses from still active workers.

At the moment, the task code cannot be reloaded after the worker has started. Thus, if a task code is changed and the jobs are being sent to the same workers, the new code will not take effect, which can be very confusing for the user. There are two possible solution to that problem. One is to version the tasks and have the workers unload the old classes and load new ones when a new version is available. This is rather tricky to achieve as there are not good provisions made for such functionality in the JVM. Another is to split the worker application into two processes—the daemon and the worker. The daemon would spawn the worker and pass it jobs it received from the master. Upon detecting that an arriving job refers to a new version of a task, the daemon would simply kill the current worker

and spawn a new one that would be a clean slate for the code.

Scala (Odersky et al., 2004) is an upcoming language that is lauded as Java successor. It runs on the JVM, and can be easily mixed with Java, but has a much more compact syntax, provisions for functional programming and deeply follows novel programming language concepts. Kaichu would gain in simplicity and expressiveness if it fully supported Scala.

Kaichu is available under the open-source Apache License, Version 2.0 and can be downloaded or forked from http://bitbucket.org/mcek/kaichu/.

Appendix A shows listings and a description of a simple Kaichu application.

Part II

Wet-lab Validation of a microRNA Library

Chapter 5

Materials and Methods

5.1 Preparation of chemically competent cells

Preparation of the 10G, *ccdB* sensitive *E. coli* cells was done using the Zymo Research Z-Competent™ *E. coli* Transformation Kit, catalogue No. T3002, in a procedure based on the instruction manual.

50 µl from a fresh, 5 ml overnight culture (LB and Streptomycin at 50 µg/ml) was added to 100 ml SOB medium with Streptomycin at 50 µg/ml in a 500 ml flask. The flask was shaken at 250 rpm at room temperature until the culture reached an absorbance between 0.4 and 0.6 at 600 nm ($OD_{600\,nm}$). The $OD_{600\,nm}$ was measured with the NanoDrop 2000c spectrophotometer using cuvettes.

The cells were pelleted by centrifugation at 3 000 rpm for 10 minutes at 4 °C. After removing the supernatant the cells were gently resuspended in 5 ml ice-cold 1X Wash Buffer supplied with the kit.

The cells were pelleted again as above and after removing the supernatant the cells were gently resuspended in 5 ml ice-cold 1X Competent Buffer supplied with the kit.

The solution was divided into 50 µl aliquots between 1.5 ml eppendorf tubes that had been frozen to −80 °C in order to flash freeze the cells. The stock was stored at −80 °C.

5.2 Oligonucleotide phosphorylation

A mix of 3 µl oligonucleotide at 100 mM (resuspended in TE 8 pH), 1 µl T4 Polynucleotide Kinase, 5 µl T4 Ligase Buffer, and H_2O to a total volume of 50 µl was incubated for 30 min at 37 °C. The enzyme was inactivated by heating to 65 °C for 20 min.

5.3 Oligonucleotide annealing

50 µl of forward strand of oligonucleotide phosphorylated as described in section 5.2 was mixed with 50 µl of reverse strand prepared in the same way. To this 2 µl of NaCl at 4M was added and topped up with Buffer TE, pH 8 to 200 µl total. The mix was heated to 94 °C for 4 min and then cooled to room temperature slowly for about 40 min.

5.4 Restriction digest

Between 600 and 800 ng of plasmid DNA was mixed with 0.5 µl of each restriction enzyme, 2 µl BSA 10x (if needed), 2 µl appropriate NEB buffer, and H_2O to a total of 20 µl. The mix was incubated for 90 min at 37 °C and the enzymes were heat inactivated for 20 min at 65 °C.

5.5 DNA extraction from an agarose gel

The gel extraction procedure was carried out using the QIAquick Gel Extraction Kit from QIAGEN according to the manufacturer's instructions with minor changes. All centrifuging steps were done for 1 minute at 13 000 rpm.

The DNA was excised from an agarose gel and placed into microcentrifuge tubes. The gel was dissolved in Buffer QG three times the volume of the gel at 50 °C in 10 minutes. The solvent was confirmed to have pH $\leqslant$ 7.5 by checking that it was yellow. The DNA in the solvent was bound to the membrane of a QIAquick column by centrifuging it. To remove the remaining traces of agarose, the column was additionally washed by centrifuging with additional 500 µl Buffer QG. The

DNA was washed with 750 µl Buffer PE, centrifuged, and centrifuged again to remove all traces of ethanol. The QIAquick column was placed into a clean microcentrifuge tube, and 8 µl Buffer EB was added of the column membrane. The column was left to stand for 1 minute and then centrifuged.

The eluted DNA was stored at −20 °C.

5.6 DNA ligation

Between 50 and 80 ng of linearised plasmid was mixed with 0.5 µl of each insert prepared as described in section 5.3, 1 µl T4 Ligase Buffer, 1 µl T4 Ligase, and H_2O to total of 10 µl. The mix was incubated for 16 h at 16 °C . The ligase was heat inactivated for 10 min at 65 °C.

5.7 *E. coli* transformation

4 µl of plasmid DNA was added to 50 µl of stock of chemically competent bacteria prepared as described in section 5.1, and gently mixed. After incubating the cells on ice for 20 minutes the bacteria were transformed by plunging the tubes into a water bath set at 42 °C for 45 seconds and then incubated on ice again for 2 minutes. 400 µl of SOB media was added to the tube and the bacteria were grown in an incubator at 37 °C for 1 h, shaking at 200 rpm. After that time the culture was added to 4 ml of LB media and Kanamycin (50 µg/ml) or Ampicillin (100 µg/ml) in a 50 ml tube, and the bacteria were left to grow for 16 h at 37 °C, shaking at 200 rpm. Alternatively, if the plasmid DNA had been freshly ligated, the cells were plated on agar plates containing the appropriate antibiotic at the same concentration as given above.

5.8 Miniprep

The minipreps were done using the QIAprep Spin Miniprep Kit (250) from QIAGEN (Cat. no. 27106). A fresh 5 ml, transformed overnight culture was centrifuged at 8 000 rpm for 10 minutes. The pellet was resuspended in 250 µl Buffer

P1 and transferred into a 1.5 ml eppendorf tube. In order to lyse the bacteria, 250 µl Buffer P2 was added and gently mixed by inverting the tube. The lysis was stopped by the addition of 350 µl Buffer N3 and mixing again. The solution was then centrifuged in a microcentrifuge for 10 minutes at 13 000 rpm. The supernatant was transferred to a QIAprep spin column and centrifuged for 60 seconds. After discarding the flow through the DNA was washed by adding 0.75 ml Buffer PE and centrifuging for 60 seconds. The flow through was discarded again and the spin column centrifuged for another 60 seconds to get rid of the residual ethanol from Buffer PE. In order to elute the DNA the column was placed in a clean 1.5 ml eppendorf tube, 50 µl Buffer EB was added to the column, left to stand for 1 minute, and then centrifuged for an another minute.

The concentration of the extracted plasmid was measured by the NanoDrop 2000c spectrophotometer at 260 nm and DNA was stored at −20 °C.

5.9 Maxiprep

A fresh 100 ml, transformed overnight culture was centrifuged for 15 minutes at 2 000 rpm. The resulting pellet was homogenously resuspended in 10 ml Buffer P1. 10 ml Buffer P2 was then added. The solution was thoroughly mixed and left at room temperature for 5 minutes for lysis to complete. The lysis was terminated by adding 10 ml Buffer N3 and thoroughly mixing again. The lysate was further incubated at room temperature for 10 minutes inside the filter cartridge and cleared by gently pushing it through the filter. The filtered lysate was mixed with 2.5 ml Buffer ER and incubated on ice for 30 minutes. During that time a QIAGEN-tip 500 was equilibrated by applying 10 ml Buffer QBT and allowed to drain through by gravity flow. The prepared tip was then used to filter the lysate, also by gravity flow. The tip was then washed twice in a similar fashion with 30 ml Buffer QC. The bound DNA was eluted from the tip resin with 15 ml Buffer QN and then precipitated by adding 10.5 ml room temperature isopropanol and mixed. In a departure from the standard protocol the elute was redistributed into 1.5 ml tubes. These were centrifuged in a Sigma 12167 centrifuge at 4 °C, 12 000 rpm for 30 minutes. The supernatant was carefully removed from each tube with a micropipette. The pellets were then washed with 40 µl ethanol each and

centrifuged again at 13 000 rpm (14 000 g) for 10 minutes at room temperature. The supernatant was gently removed again and the pellets were left to dry on air for 10 minutes. Finally, the pellets were redissolved in 30 µl ddH_2O each and solutions combined.

5.10 Gateway LR reaction

The LR reaction mix was prepared in a slight departure from the manufacturer's instructions in order to economise on the LR Clonase.

5 fmol/µl of each entry vector (one containing the promoter and the other containing the gene) were mixed with 10 fmol/µl of the appropriate donor vector, 0.5 µl LR Clonase II, and TE pH 8 to a total of 5 µl. The mix was incubated for 16 h and after that 0.5 µl Proteinase K was added and incubated for 10 min at 37 °C.

5.11 HEK 293 cell culture

HEK293FT (Invitrogen) cells were maintained in Dulbecco's modified Eagle medium (DMEM, Cellgro) supplemented with 10% FBS (Atlanta biologicals), 1% penicillin/streptomycin (Cellgro) and non-essential amino acids (HyClone) at 37 °C, 100% humidity and 5% CO_2.

5.12 HEK 293 cell transfection

All transfections were carried out using attractene transfection reagent (Qiagen #301005) according to manufacturer's protocol with minor modifications. A total of 200 ng of DNA was diluted in base DMEM (Cellgro #10-013) to a final volume of 60 µl. Then, 1.5 µl of attractene was added, the mix was vortexed briefly and incubated at room temperature for 15 minutes. Cells were trypsinized, counted and seeded in 24-well plates (2×10^5 cells per well in 500 µl of complete medium). Shortly after seeding, the transfection complexes were added to the cell suspension and mixed (by rocking the plate). Plates were incubated in standard cell culture

conditions (humidified, 37 °C, 5% CO_2 incubator, Section 5.11). The next day cells were supplemented with 1 ml of fresh complete media. Fluorescence was assayed 48 h post transfection using fluorescence microscopy and flow cytometry.

5.13 Flow cytometry

Cells were analyzed with LSRFortessa flow cytometer, equipped with 405, 488 and 561 nm lasers (BD Biosciences). 30,000 events per sample were collected, using a forward scatter threshold of 5,000. Fluorescence data were acquired with the following cytometer settings: 561 nm laser and 610/20 nm filter for mKate and 405 laser and 525/50 filter for Cerulean. Data analysis was performed with FACSDiva software (BD Biosciences) and Microsoft Excel and Matlab.

Chapter 6

Experimental Design

6.1 Introduction

The goal of this work was to design and implement a method for discovery of ensembles of artificial miRNAs that would be: efficient, orthogonal with respect to the host system, and orthogonal with respect to each other in the ensemble.

As part of this project, an ensemble of eight miRNAs (MT1-8) was generated in the context of the human host. The experiments described in this chapter aimed to test those sequences against the given requirements in the wet lab.

However, direct test for the host orthogonality requirement was not feasible in this project as a true assessment of the global effect of an miRNA on the system would have to involve a measurement of all protein levels. Microarray or transcriptome sequencing measurements would have been inadequate as miRNAs often do not affect the transcript levels. The pSILAC method (Selbach et al., 2008) that allows measurements of protein levels was still out of reach financially.

In order to asses whether a multiplicity of weak off-targets significantly diminishes the efficiency of miRNAs, two sequences (OT1-2) were generated such that they had many predicted off-targets while scoring well against the siRNA rules.

Further sections explain the technical details of the experiments, Section 6.3 describes the experimental setup for testing the efficiency, Section 6.4 describes the experiment for testing the inter-ensemble orthogonality.

The experimental part of the project was carried out in conjunction with Dr Liliana Wróblewska from the Ron Weiss lab at the Massachusetts Institute of Technology (MIT) in USA. The final experimental set up was designed with the lab's know how and all of the transfection and FACS experiments were done at MIT while most of the construction and bacterial cloning of the vectors was done locally at Brunel University.

6.2 Overview of the experimental process

The sequences designed in Chapter 3 were tested in circuits that comprised two parts—the miRNA expression construct and the target construct. Both were cloned into separate Gateway destination vectors.

The miRNA expression construct (Figure 6.1A) expressed the far-red fluorescent monomeric Katushka (mKate) protein (Shcherbo et al., 2007) with an intronic pri-miRNA. The red colour would be used to confirm miRNA expression. The target construct (Figure 6.1B) expressed the cyan fluorescent Cerulean protein (Malo et al., 2007) containing an miRNA target site in the 3' UTR. Cyan colour would be used to assess the level of repression by an miRNA. Expression in both constructs was driven by the human constitutive hEF1a promoter. The choice of cyan and red colours was motivated by the fact that they are relatively far from each other in the electromagnetic spectrum which reduced the potential crossover between them.

6.2.1 Subcloning strategy

For reasons of reusability the Gateway cloning system (Hartley et al., 2000) was used. In this framework, the genes and promoters are kept separately in the so-called entry vectors (pENTR). Those can be treated as library vectors where the parts are kept for easy reuse and bacterial amplification. The entry vectors are built by standard recombination methods and can be amplified in kanamycin sensitive *E. coli*. The vectors that are ready to be transfected into mammalian cells are called destination vectors (pEXP) and are built in an LR reaction from an entry vector containing a promoter, an entry vector containing a gene, and a

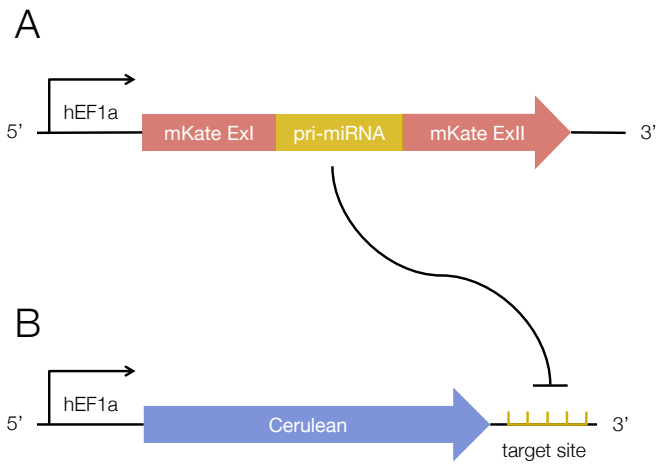

Figure 6.1: **The test circuit. (A)** The miRNA expression construct constitutively expressed a far-red fluorescent mKate protein with an intronic pri-miRNA. **(B)** The target construct consitutively expressed cyan fluorescent Cerulean protein with a target site in its 3' UTR. A matching mature miRNA would repress the translation of the Cerulean protein.

donor vector. Figure 6.2 explains this process in more detail.

In the LR reaction the DNA is recombined such that the promoter and the gene are subcloned into the donor vector which serves as the backbone of the future destination vector. In this process, a lethal *ccdB* gene (Bernard and Couturier, 1992) is cut out from the donor vector and replaced with the inserts. The mix can be then transformed into *ccdB*-sensitive *E. coli*. This setup ensures that the donor vectors that have not been processed would not propagate in a culture. In order to select for the transformed cells, the donor vector carries the Ampicillin resistance gene and the growth media is supplemented with Ampicillin.

6.2.2 MicroRNA expression

As mentioned in Section 2.4, miRNAs can be produced from transcripts that fold into hairpins or hairpin-containing structures. Given a proven pri-miRNA DNA coding sequence it is possible to replace the passenger and guide strand portions to prepare expression of an arbitrary miRNA.

In this project an expression system earlier successfully used in the Weiss lab

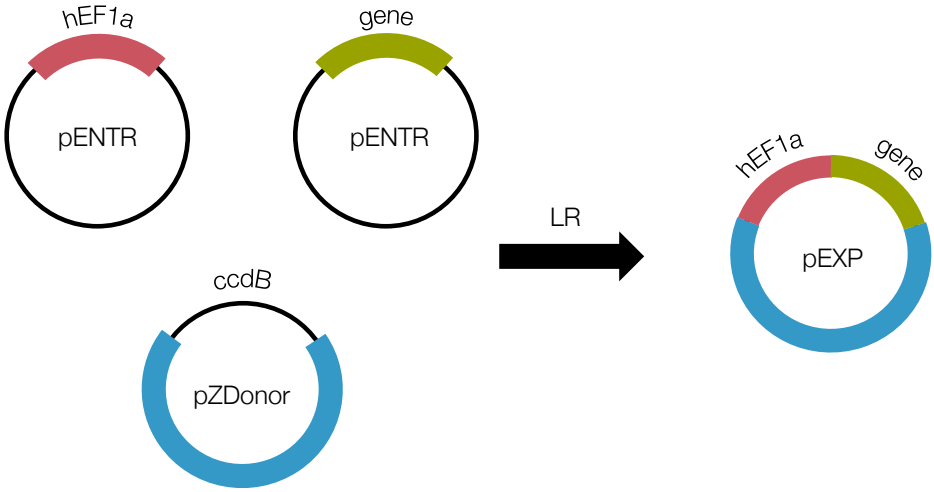

Figure 6.2: **The LR reaction performed to make the transfection vectors.** The pZDonor vector is a backbone for the future expression vector, but it also expresses the lethal *ccdB* gene, so cells transfected with the unmodified pZDonor vector will not survive. The two pENTR vectors contain the promoter (hEF1a) and the gene aimed for expression (either the mKate+miRNA or the Cerulean+target site construct). In a one-step LR reaction the ccdB gene is excised from the pZDonor vector and is replaced with the promoter and the gene from the pENTR vectors in the correct order, so that the resulting pEXP vector can be transformed into human cells and induce expression of the gene.

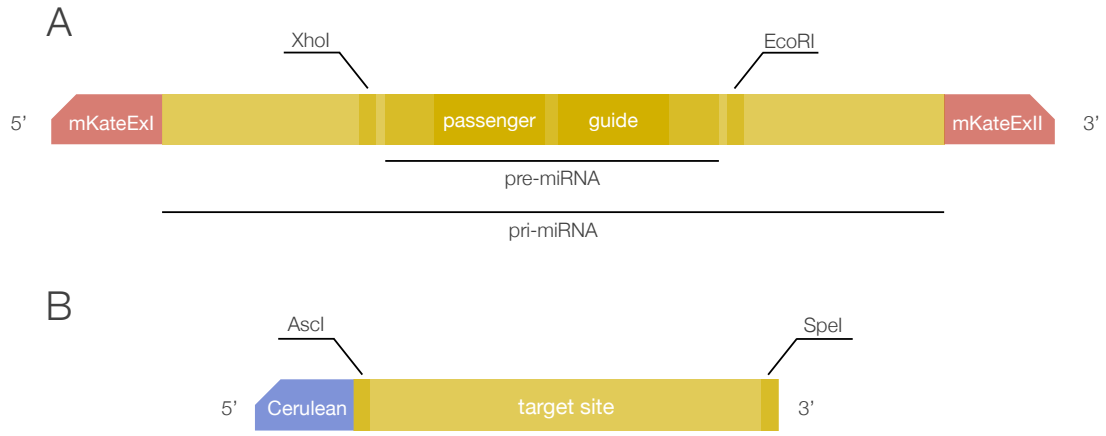

Figure 6.3: **Structure of the expression constructs. (A)** The pri-miRNA is situated in an intron of the mKate protein gene. The pre-miRNA, containing the passenger and guide strands, is surrounded by XhoI and EcorI restriction sites. **(B)** The target site is placed within the 3' UTR of the Cerulean gene (encodes cyan fluorescence) and is surrounded by the AscI and SpeI restriction sites. Such design allows for easy replacement of the functional parts of the miRNA:target site pair in the expression cricuit.

(Rinaudo et al., 2007*b*; Xie et al., 2011) was utilised. In this template, a 347 pri-miRNA is expressed from an intron of the mKate gene (encodes red fluorescence). The pre-miRNA portion is enclosed by XhoI and EcoRI restriction sites that allow the miRNA sequence easily replaced (Figure 6.3A). The construct was stored in a pENTR vector.

6.2.3 Target construct

For each of the 8+2 generated miRNAs a target construct was made, so that the efficiency of repression by the respective miRNA and potential off-targeting by other miRNAs could be measured (Also used in Xie et al. (2011) and Rinaudo et al. (2007*b*)). In this construct, a cyan Cerulean protein would be fashioned to harbour a single target site in its 3' UTR.

Again, a template pENTR vector was used and the target site was surrounded with AscI and SpeI restriction sites, so that the vector could be linearised and ligated with a new target site insert (Figure 6.3B).

6.2.4 Construction of the miRNA expression and target site construct plasmids

The inserts required to build the constructs were relatively short. The target site insert consisted only of the ligation overhangs required to combine it with an expression plasmid and the target site itself—29 nt. The miRNA expression insert, apart from the ligation overhangs, comprised the whole pre-miRNA hairpin to include both the passenger and the guide strand—103 nt. Those could be easily constructed from single-stranded oligonucleotides synthesised by an external company.

The target site inserts could be built out of two single-stranded oligonucleotides. However, 103 nt long single-stranded oligonucleotides are difficult to achieve because the errors in synthesis diminish the rate of correct sequences as the length of the sequences increases. Therefore, such sequences are not available commercially, and the miRNA expression construct inserts were built from four rather than two oligonucleotides.

A simple backtracking algorithm was implemented to generate oligonucleotides from an arbitrary sequence, such that: the length of the oligonucleotides was not longer than 60 nt, and the oligonucleotides would anneal into short inserts with overhangs such that they would ligate in the correct order to form the desired double stranded DNA. The algorithm uses 4 nt internal overhangs and ensures at least 2 nt of non-complementarity between the overhangs that should not match to make the ligation unambiguous.

To summarise the construction of the expression plasmids, the single-stranded oligonucleotides were phosphorylated to aid later ligation (Section 5.2), and the complementary pairs annealed (Section 5.3); the template vectors were linearised (Sections 5.4, 5.5), and eventually ligated with the insert(s) (Section 5.6).

The prepared mix was transformed into competent *E. coli* (Section 5.7), plated on an agarose gel, and grown in LB broth supplemented with kanamycin (50 µg/ml). The extracted plasmid DNA sequence (Sections 5.8, **??**) was confirmed by synthesis at an external company.

The confirmed pENTR vector, the hEF1a pENTR vector, and a Gateway donor vector were used in the LR reaction (Section 5.10) to form the pEXP

vector. The resulting mix was transformed into *ccdB*-sensitive competent *E. coli*, plated, and grown in LB broth supplemented with ampicillin (100 µg/ml). After extraction, the plasmid DNA sequence was confirmed by sequencing at an external company.

The above work resulted in 20 pENTR library vectors—10 containing the miRNA expression genes and 10 containing the target site expression genes—for each of the MT1-8 and OT1-2 miRNA sequences designed in Chapter 3. The LR reactions resulted in another 20 respective pEXP vectors. Those were ready constructs with the hEF1a promoters that could be transfected into human cells.

The above work was done at Brunel by the author and all further work—on transfection into the human cells, flow cytometry readouts and repeat of the MT2 Cerulean and paris for MT6 and MT8 alternatives (described in Section 7.3.2 in the next chapter)—was done by Dr Liliana Wróblewska at MIT.

6.3 Efficiency of artificial miRNAs

In order to assess the efficiency of the 8+2 miRNAs, the HEK 293 cells were cotransfected with the complementary expression and target constructs (Section 5.12). 100 ng of the reporter vector pExpr-hEF1a-Cerulean-MTx, and 100 ng of vector expressing mKate fluorescent protein and intronically encoded microRNA, pExpr-hEF1a-mKateEx-miRMTx were used. The results were measured by fluorescent microscopy and flow cytometry 5.13.

A negative control for each sequence was a cotransfection of the target site construct with an mKate expression construct that lacked the miRNA intron altogether (pExpr-hEF1a-mKate). The cyan produced by the negative control would provide a baseline for assessment of repression strength.

The experimental design and the predicted results are illustrated in Figure 6.4.

6.4 Inter-ensemble orthogonality

The second orthogonality requirement (Section 3.2.3) stated that the members of an ensemble should not cross target with each other. In order to check for that

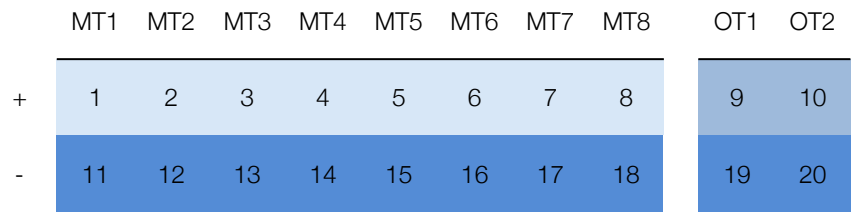

Figure 6.4: **Experimental design for measuring the efficiency of the designed sequences.** The test samples 1-8 represent the predicted results of co-expression of matching mkMTX and CMTX constructs. The predicted Cerulean (cyan) output should be significantly lower than in the case of the negative control (samples 11-18) where the CMTX construct was cotransfected with mKate protein without the miRNA intron. The off-targeting OT1-2 miRNAs might have reduced efficiency due to a large number of predicted off-targets (samples 9-10).

every combination of miRNA:target site in the ensemble has to be tested. The cotransfection matrix with sample numbering and expected results is presented in Figure 6.5. It is hypothesised that matching miRNA:target site pairs would result in low expression of the Cerulean protein, and mismatching pairs will not affect the production of Cerulean which would be as high as in the negative control.

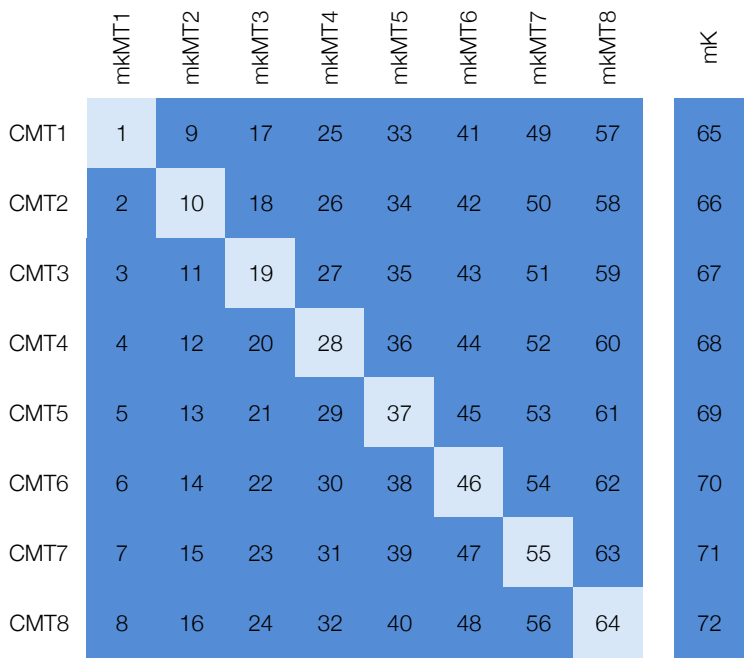

Figure 6.5: **The cotransfection matrix.** Each square represents a cotransfection experiment. The columns represent miRNA expression constructs (mkMT1-8) and the negative control (mK). The negative control construct expresses the mKate protein without the miRNA intron. The rows represent the target site constructs (CMT1-8). Each sample is numbered as shown. The table also shows the predicted results. Ideally, the matching pairs of miRNA:target site would result in no or reduced expression of the Cerulean protein (pale blue) compared to the negative control indicating high efficiency as in Section 6.3. Meanwhile, the negative control would be predicted to express Cerulean as should the samples with mismatching pairs (blue). This would indicate no cross-talk between the miRNAs in the ensemble.

Chapter 7

Results

7.1 Introduction

The aim of this project was to develop a computational framework for design of efficient and orthogonal synthetic miRNA sequences. As part of it, a specific set of eight such sequences for use in human cells was produced and two additional sequences that complied with the state of the art siRNA design rules, but were not orthogonal with respect to the human transcriptome.

The sequences were designed as described in Chapter 3. Chapter 6 set forth the experimental design of experiments to validate those sequences against the requirements of efficiency and orthogonality. In this chapter results of those experiments are described and discussed.

The chapter starts by mentioning the early batch of the miRNA sequences that were 21 nt long, how their validation failed, and the identification of the cause for it. Then it moves to report the results based on the second batch of the sequences.

The lab experiments presented in this chapter were performed by Dr Liliana Wróblewska from the Ron Weiss lab at the Massachusetts Institute of Technology (MIT). The preceding plasmid construction and subsequent data analysis was done by the author at the Brunel University.

Table 7.1: The eight 21 nt long miRNA candidates originally selected using the Reynolds rules (Section 2.6). The sequences are the miRNA guide strands. The seeds are in boldface.

Name	Sequence
$MT1_R$	A**TTACGCG**AATACCGGTCGTT
$MT2_R$	T**GTTGGGT**AAATAACCTAATT
$MT3_R$	A**TAGGGTT**AACCGCCCTAATT
$MT4_R$	A**TTTCGTG**GAAACGCCTAATT
$MT5_R$	A**GAGTGGG**TACCCTAATAATT
$MT6_R$	A**CGTTTGT**AACCCGCCTAATT
$MT7_R$	T**TTGGCGG**TATACCGCTAATT
$MT8_R$	T**GTGTTTC**GACGCGCCTAATT

7.2 21 nt long miRNAs (first batch)

Chapter 3 describes the final process in which the miRNA sequences were designed, but initially a slightly different process was used that used the Reynolds siRNA design rules (Section 2.6) and produced 21 nt rather than 22 nt long sequences.

The resulting sequences (Table 7.1) were tested according to the experimental design described in the previous chapter. However, none of the matching miRNA:target site pairs performed as expected—cotransfection of circuits containing matching miRNA:target site pairs resulted in Cerulean expression very close to the negative controls where no miRNA was used.

It was hypothesised that using a miRNA expression template originally intended for 22 nt long miRNAs was the cause. Zeng and Cullen (2003) performed exhaustive tests on how the secondary structure of the pri-miRNA can affect the processing in the miRNA biogenesis. One of the cases in which the processing was markedly hindered was when a bulge existed close to the Drosha cleavage site. Indeed, such a bulge exists in the template used and a shorter miRNA sequence would have brought it closer to the cleavage site (Figure 7.1).

To further test this hypothesis, two alternative sequences were designed by prepending the 'A' nucleotide to miRNAs $MT1_R$ and $MT2_R$ forming AATTACGCGAATACCGGTCGTT and ATGTTGGGTAAATAACCTAATT respectively. These were tested in single repeats and the flow cytometry readings showed around

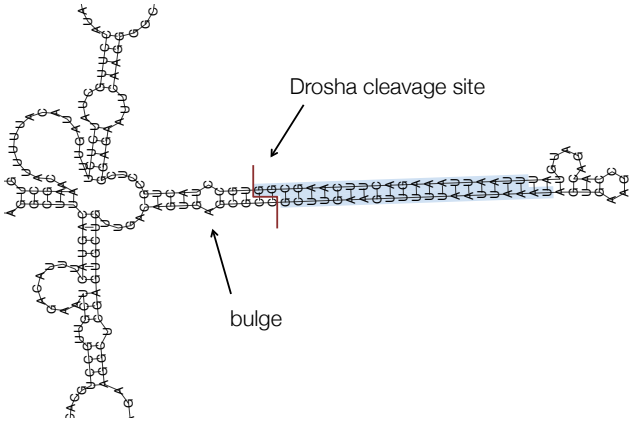

Figure 7.1: **The pri-miRNA template contains a bulge close to the Drosha cleavage site.** The figure represents a predicted secondary structure (RNAfold, Vienna RNA) of a part of the pri-miRNA produced from the miRNA expression template used in the experiments. In this case the template contains a 22 nt long miRNA sequence (the duplex highlighted in blue) and the RNA is processed efficiently. The hypothesis put forward here was that a 21 nt long miRNA sequence would bring the bulge closer to the Drosha cleavage site which would in turn disturb the Drosha action.

94% repression of the Cerulean signal for MT1$_R$' and around 78.3% repression for MT2$_R$'.

These results prompted a redesign of the computational framework to produce 22 nt long sequences. An opportunity was also taken to replace the Reynolds siRNA rules with more accurate framework that was elucidated from more extensive dataset than Reynolds had access to and based on support vector machines as described in Chapter 3.

7.3 Final batch of 22 nt long sequences

After redesigning the computational process a new ensemble of eight sequences was produced as shown in Table 3.6 in Chapter 3. Those were then tested in the wet lab as described in Chapter 6 and the results are presented below.

7.3.1 Preliminary results

The first run of the experiments tested only the matching pairs of miRNA:target site and was done in one repeat.

The Cerulean and mKate constructs containing matching target sites and miRNA expression introns were cotransfected into HEK 293 cells for each of the designed MT1-MT8 sequences. Negative controls were performed by cotransfecting each of the Cerulean construct with the mKate construct lacking the miRNA expression intron. The results were obtained by measuring the Cyan fluorescence with flow cytometry. Also, fluorescence in the Red channel was measured to confirm expression of the mKate construct.

In flow cytometry (Section 5.13) a stream of particles passing through a readout apparatus is illuminated with lasers of different wave length and fluorescence measurements in different channels of the electromagnetic spectrum is performed. A readout of a single particle is an event. In order to obtain meaningful results those events have to be filtered. First, the particles are gated on their size using the side scatter (SSC) and forward scatter (FSC). This is to ensure that the particles taken into consideration further are indeed cells and not cell debris, etc.

The mKate protein will emit light around 633 nm which fits in the Texas Red channel while the Cerulean emits light around 475 nm which fits in the AmCyan channel. In order to select for the cells successfully transfected with the mKate expression construct, the events are gated further to have a measured fluorescence in the Texas Red channel above a certain theshold. The same is done for the Cerulean construct in the AmCyan channel. The cells of interest are represented by the events at the intersection of the two sets above—those that have been successfully cotransfected.

For conciseness, the detailed flow cytomtetry results for all samples are not presented in the main body of the text. Figure 7.2 shows results from the MT1 sequence test. The graphs show the relationship between the red and cyan fluorescence and the distributions of each in the cell population. In this example there is a noticeable decrease of cyan fluorescence in the sample where the MT1 miRNA was expressed. The data carried for further analysis are the means of the fluorescence readouts from the relevant gates.

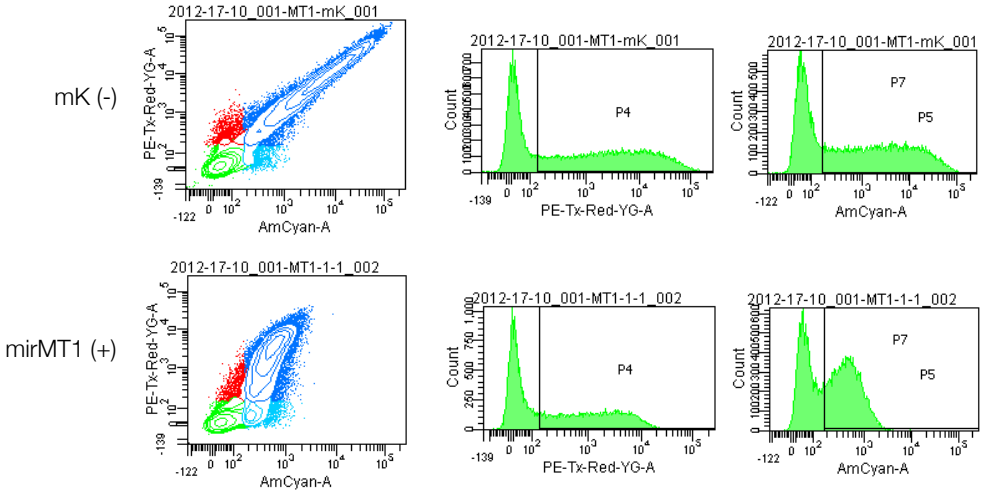

Figure 7.2: **Flow cytometry data shows repression of Cerulean production in the MT1 miRNA:target site pair.** The cells were cotransfected with the mKate and Cerulean expression constructs. The Cerulean gene in each case was tagged with the MT1 target site. The mKate gene had no intron in the negative control (top) while in the test sample (bottom) it contained an intron expressing the MT1 miRNA. Cells successfully transfected with the mKate construct are represented by the P4 gate and shown in red. Cells successfully transfected with the Cerulean construct are represented by the P5 gate and shown in light blue. The P7 gate and dark blue colour represent cells at the intersection of P4 and P5. The negative control differes from the test sample in two ways. The first is that the mKate expression is higher in the negative control cells—it goes up to 1E+5 while in the test sample only up to around 1.1E+4 (middle). This phenomenon is explained in the caption of Figure 7.6. More importantly, on the left and right graphs, it is clearly visible that the cells tend to emit less Cyan light in the test sample than in the negative sample. The means of the red fluorescence in gate P4 are 11,557 and 2,940 in the negative control and the test sample respectively while the means of the cyan fluorescence in gate P7 are 9,607 and 635 in the negative control and the test sample respectively.

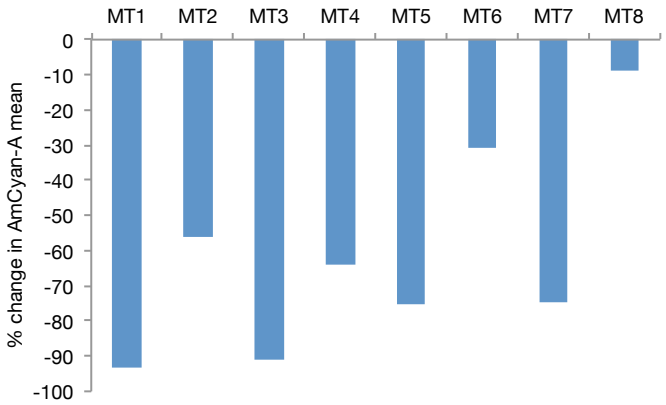

Figure 7.3: The preliminary experiments found the efficiency of MT6 and MT8 miRNAs below 50%. The figure shows the percent change in the Cerulean expression of the target construct as affected by expression of a matching miRNA. The percent change was calculated with respect to the negative controls where no miRNA was expressed.

The preliminary results suggested that most of the miRNAs of the ensemble performed well, however MT6 and MT8 caused only around 30% and 9% repression respectively (Figure 7.3). The percent change in each case was calculated with respect to the relevant negative control.

Another problem surfaced in the negative control data where the Cerulean construct appeared to be inefficient at expressing the Cerulean protein. Figure 7.4 shows the Cerulean expression levels across all Cerulean+target site constructs. The expression of the CMT2 Cerulean stood out as much lower than in any of the other negative controls. In order to explain that, the MIT lab ordered a resequencing of the pEXP CMT2 plasmid. A mutation was found in the hEF1a promoter which drives the expression of the Cerulean protein. This mutation likely debilitated the functioning of the promoter. This plasmid was then constructed again and its sequence confirmed.

7.3.2 MT6' and MT8'

As mentioned in the previous section, the two miRNA sequences—MT6 and MT8—did not perform as well as expected in the preliminary experiments. Those

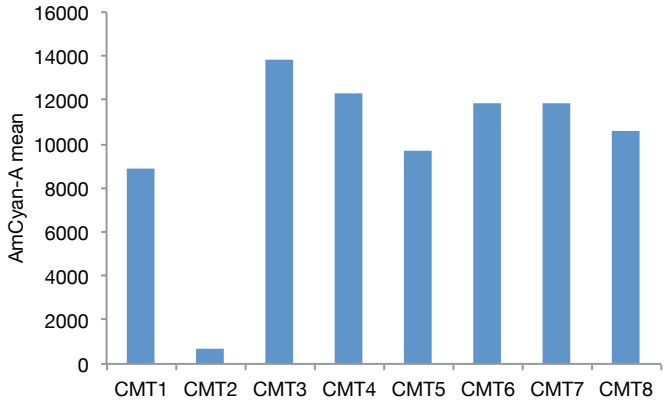

Figure 7.4: In the preliminary experiments the expression of Cerulean by the CMT2 was found markedly lower than in others. The figure presents the mean of raw fluorescence flow cytometry readouts for the negative controls—where the Cerulean constructs are expressed, but the mKate constructs lack the miRNA.

two were replaced with their alternatives as follows.

MT6 was replaced with one of the 19 alternatives from the 20 final MT6 variants of the computational design process (Section 3.8). It was chosen to be one of the best in terms of orthogonality with respect of the well-performing miRNAs in the ensemble and it had two additional 'C' nucleotides close to the 3' side to improve the free energy imbalance of the duplex. This is important for the preferential guide strand choice and should shift the balance towards the intended guide strand.

The MT8 sequence, in spite of its high score assigned by the SVN siRNA rules algorithm, had unfavourable free energy imbalance of the duplex which was mostly caused by its seed being very G/C rich. This was allowed during the seed filtering stage (Section 3.4.2) where only at two out of five preferred positions in the seed 'A' or 'T' were required. Setting a stricter requirement at that stage should result in production of better miRNA sequences.

In order to improve the duplex free energy imbalance without changing the sequence dramatically, the seed was replaced with another seed that: (1) was only different by one nucleotide, replaced a G/C into an A/T, (2) scored well in the seed ranking (Section 3.4.1). The chosen MT6' and MT8' sequences are shown in Figure 7.5 together with MT6 and MT8 for comparison.

```
MT6   5' TCAACAAGCCATCCACAAACAT 3'
MT6' 5' TCAACAAGCCATACAACCCTA 3'

MT8   5' TACCCGATAAAACACAAACCTA 3'
MT8' 5' TAACCGATAAAACACAAACCTA 3'
```

Figure 7.5: The inefficient MT6 and MT8 miRNA sequences were replaced with similar alternatives in an attempt to improve their duplex free energy imbalance to aid the preferrential choice of the intended guide strand during their biogenesis.

7.4 Efficiency

The final set of miRNAs was assessed for efficiency as described in Chapter 6. The HEK 293 cells were cotransfected with the mKate+miRNA construct and the matching Cerulean+target site construct. First, images in the Texas Red and Cyan channels were taken under a microscope. Figure 7.6 represents the results obtained with the MT1 sequence. The Texas Red channel on the right pictures cells that have been successfully transfected with the mKate+miRNA construct, or just mKate in the case of the negative control. The Cyan channel (left) pictures expression of the Cerulean protein. Cyan cells are clearly visible in the negative control (top left) while very little expression is visible in the sample where the miRNA was expressed indicating a high level of repression.

Figure 7.7 shows the microscope imagery for all eight sequences of the ensemble in the Cyan channel and allow for a visual assessment of the repression efficiency.

The same cells were then trypsinised and the fluorescence was measured with flow cytometry. Figure 7.8 shows the means of raw fluorescence flow cytometry measurements of the samples and negative controls as well as fluorescence that resulted in transfection with a Cerulean construct not containing any target site (Cer-1).

It is important to note that the Cerulean output with the gene tagged with a target site is already consistently lower than the Cerulean output with the gene lacking the target site before any miRNA is applied. In this case the percent decrease ranges between 27 and 51% and the differences between the CMT1-8 and Cer-1 are all statistically significant according to a single-tailed t-test that

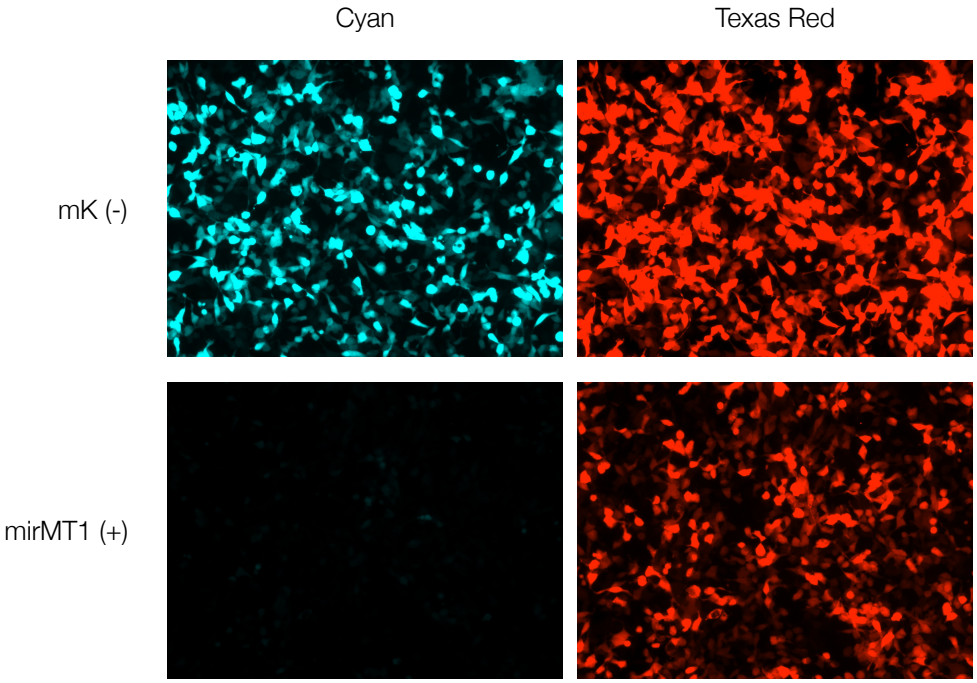

Figure 7.6: **Fluorescent microscopy images of the MT1 sequence transfection shows a marked decrease of Cerulean expression when miRNA is present.** Images show the transfected HEK 293 cells visible in two colour channels—the Cyan (left) and the Texas Red (right). The first row represents the negative control where the expressed mKate protein does not contain the miRNA intron. The results show marked expression of the Cerulean protein in the cyan channel and expression of the mKate protein in the red channel. The second row shows the test sample where the MT1 miRNA is expressed. The Cerulean protein is hardly visible. The expression of mKate is noticeably lower than in the negative control (typical for all samples MT1-8). This can be explained by the existence of the miRNA intron and thus necessity of splicing which in turn lowers the overall yield of mKate. This does not have any bearing on the results however.

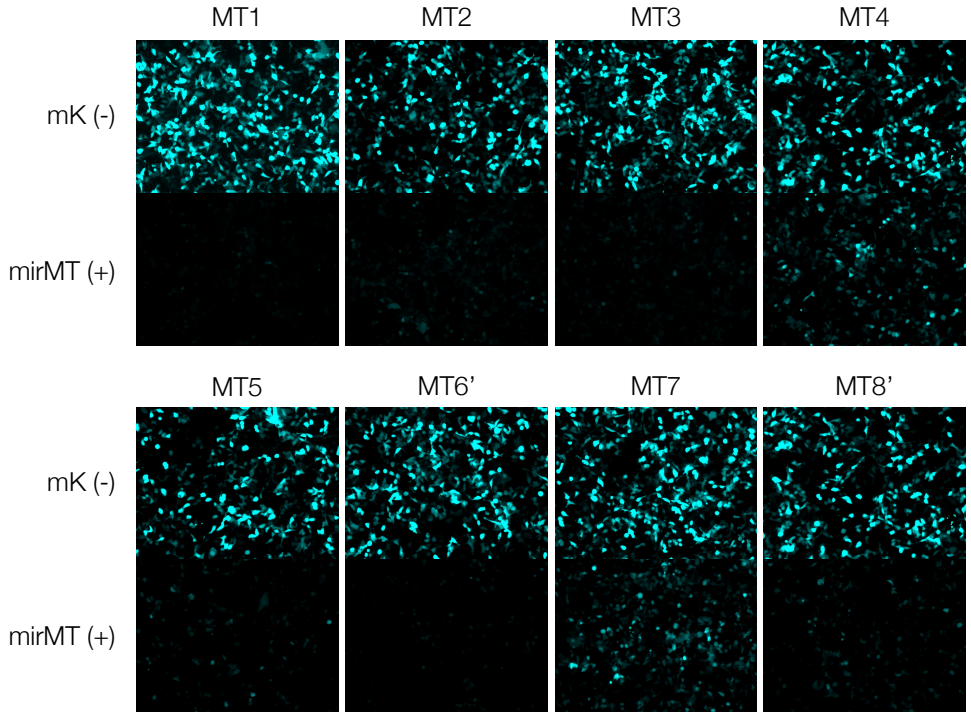

Figure 7.7: **Fluorescent microscopy images of the MT1-8 transfection shows miRNA repression of the Cerulean gene for all test samples.** The images were taken before the flow cytometry readouts and show the cyan channel for the Cerulean construct cotransfected with the matching miRNA mKate construct (bottom). The negative controls, where each of the Cerulean constructs were cotransfected with the mKate construct lacking the miRNA intron, are shown on the top.

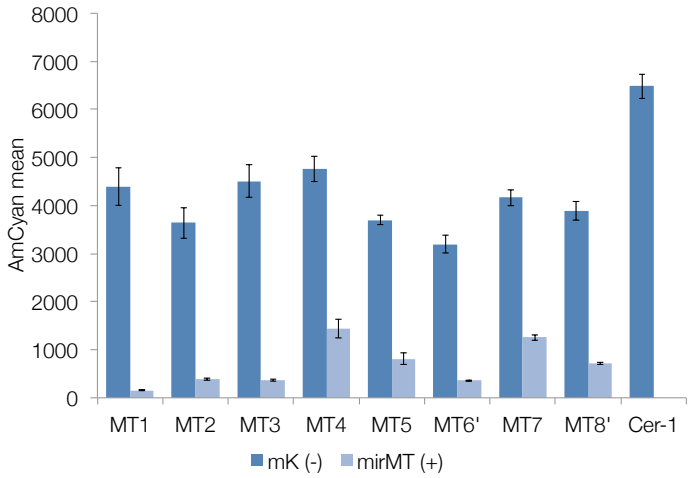

Figure 7.8: **Repression by miRNAs MT1-8 results in a marked decrease in expression.** The figure presents means and standard deviations of Cerulean fluorescence readouts using flow cytometry. The negative controls are shown in dark blue and the matching cotransfection samples are shown in light blue. Additionally, a control of a single transfection of the Cerulean gene lacking any target site is shown (Cer-1). The decrease in expression between the matching cotransfection experimental samples and negative controls is statistically significant. The P-values were calculated using a single-tailed t-test without assuming equal variance. All P-values are <0.0015.

does not assume equal variance with P-values all below 0.001. This effect should be taken into consideration in applications.

A single-tailed t-test was performed for each of the samples to assess the significance of the decrease with respect to the negative controls. Equal variance was not assumed. All calculated P-values were below 0.0015 which made the decreases highly significant.

Figure 7.9 shows the percent change in expression caused by the miRNAs on their intended targets. The final sequences are between 70 and 96% efficient. In applications, more than one target site can be put on a target to further improve the repression. Variable repression strengths may however be useful to fine tune a synthetic circuit and these results can serve as a guide to the relative efficiency of the miRNAs.

The following results expand into testing the orthogonality of the miRNAs

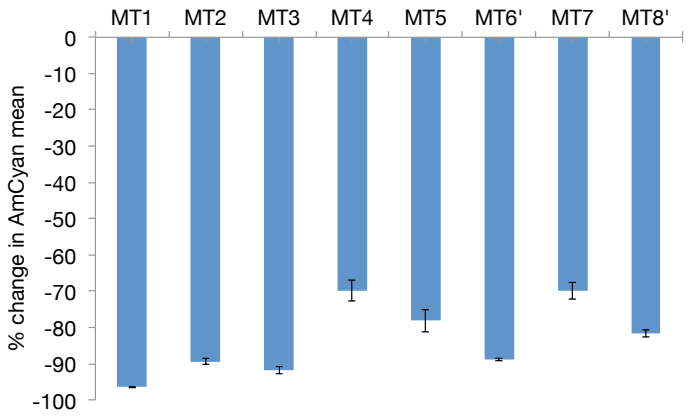

Figure 7.9: **MiRNAs MT1-8 cause at least 70% repression.** The figure shows the same data as Figure 7.8, but with each sample normalised with respect to its negative control to show percent change in repression for better comparison between the efficiencies. The error bars show standard deviation.

within the ensemble.

7.5 Inter-ensemble orthogonality

In order to assess the inter-ensemble orthogonality, all combinations of the target site (CMT1-8) and the miRNA expression (kMT1-8) constructs were cotransfected. The aim was to confirm that there was no cross-talk between the miRNAs and that they can be used in a single system and working separation.

Three experimental replicates of the cotransfection matrix were performed. The raw results from the flow cytometry read out are shown numerically and as heat maps in Figure 7.10. Each square in the heat maps 1-3 represents an experimental sample where a miRNA circuit was cotransfected with a target site construct. The intensity of the blue colour on the heat maps relates to the measured cyan fluorescence levels which in turn is a proxy for the Cerulean protein expression levels.

The samples on the diagonal of the heat maps represent the matched miRNA:target site pairs and have been analysed already in the previous section. The samples outside the diagonal represent mismatched pairs and can point to potential cross-talk between the miRNAs in the ensemble. Visual inspection of the heat maps

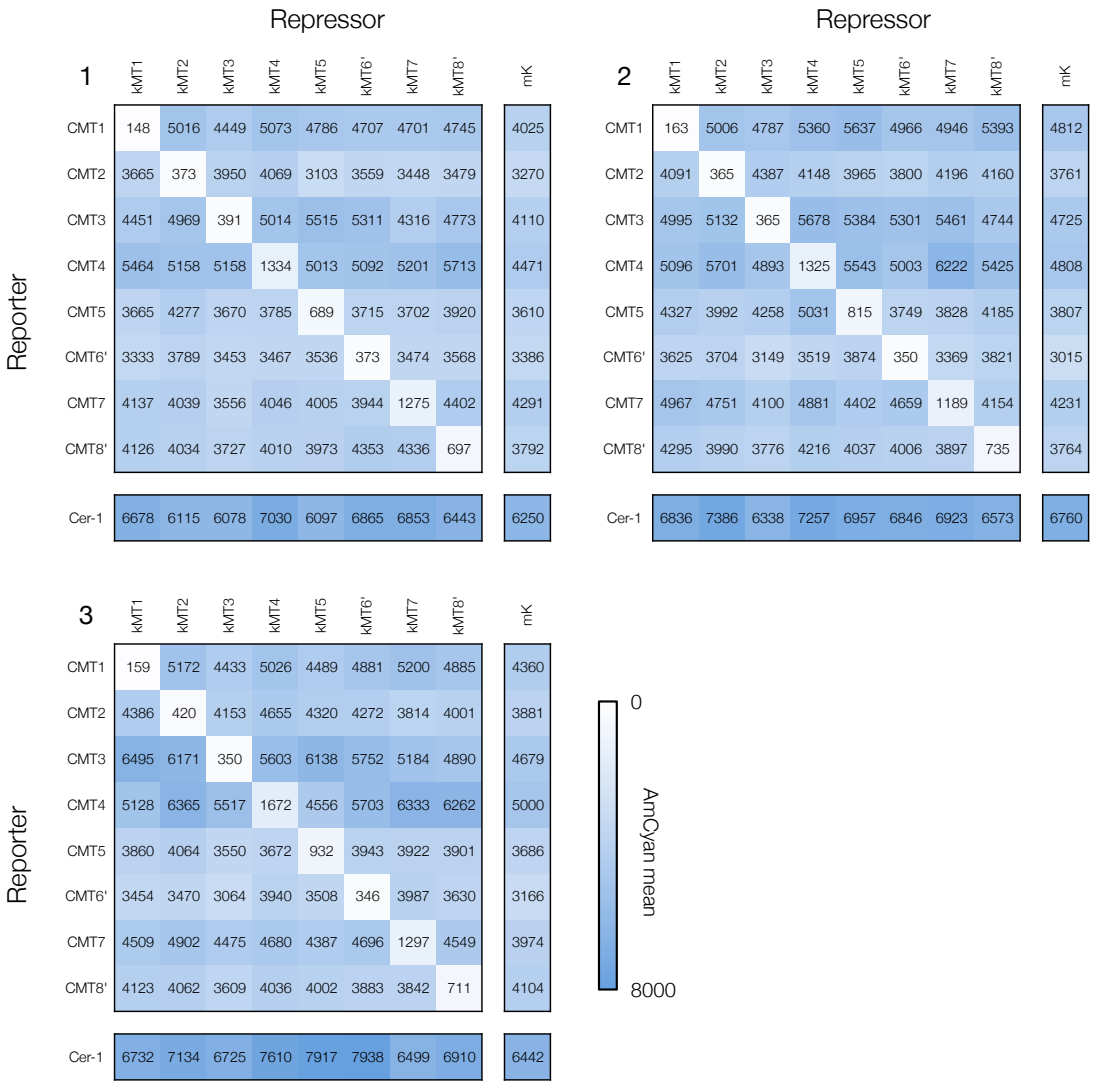

Figure 7.10: **Mismatched miRNA:target site pairs result in Cerulean expression similar to the negative controls.** The heatmaps 1-3 show results of cotransfection of the Cerulean target site constructs and the mKate+miRNA constructs. The readouts are the mean values of intensity on the AmCyan channel and are a proxy for the Cerulean protein levels. The squares show cotransfection of the constructs containing the designed sequences (kMT1-8 and CMT1-8). The columns denoted mK show experimental samples where the mKate protein had no miRNA intron. Those were the negative controls and are used for calculating the statistical significance and later for normalisation (Figure 7.11). The rows denoted as Cer-1 show samples where the Cerulean construct was transfected without any target site.

suggests that the mismatched samples are not noticeably different from the negative samples in terms of the Cerulean expression.

It is clearer to analyse the significance of these results when they are presented as a percent change in expression. Figure 7.11 shows the percent change in the AmCyan intensity as normalised against the negative controls (mK) for all of the three repeats as well as their means.

The notion that the miRNAs do not repress targets with mismatched target sites can be statistically represented as follows. The null hypothesis is that the samples represent a repression of at least δ percent for a δ that is chosen to represent a small enough repression to be accepted as irrelevant. Here, $\delta = 10\%$ was chosen. In order to decide whether the null hypothesis can be rejected, a t-statistic is calculated

$$t = \frac{\mu - (-\delta)}{\sigma/\sqrt{N}} \tag{7.1}$$

where μ is the sample percent change mean, σ is the standard deviation, and N is the sample size (here $N = 3$). The t values for each samples were used to calculate P-values for a single-tailed test with 2 degrees of freedom. The P-values are presented in Figure 7.12. Low number of repeats and relatively high variation in the results in the mismatched samples (Figure 7.13) contributed to the fact that not in all cases the null hypothesis could be rejected even though the results at face value do not suggest any marked repression occurring.

Figure 7.13 indeed suggests that the unrepressed samples tend to show increased Cerulean production comparing to the negative control. This may be caused by the fact that lower production of mKate with intronic miRNA releases more cellular resources to produce Cerulean.

The choice of accepted δ depends on the context. For instance, in applications where potential cross-talk repression of at most 25% would be acceptable, all but the MT3-MT7 pair (P-value = 0.06) would be statistically guaranteed to fulfil the requirement.

It is possible to invert this analysis and try to test whether there is a detectable repression in any of the cases that are meant to be orthogonal. The one-way ANOVA test was performed between groups for each of the target sites. The first row of the Table 7.2 shows P-values calculated for each target site, taking out the

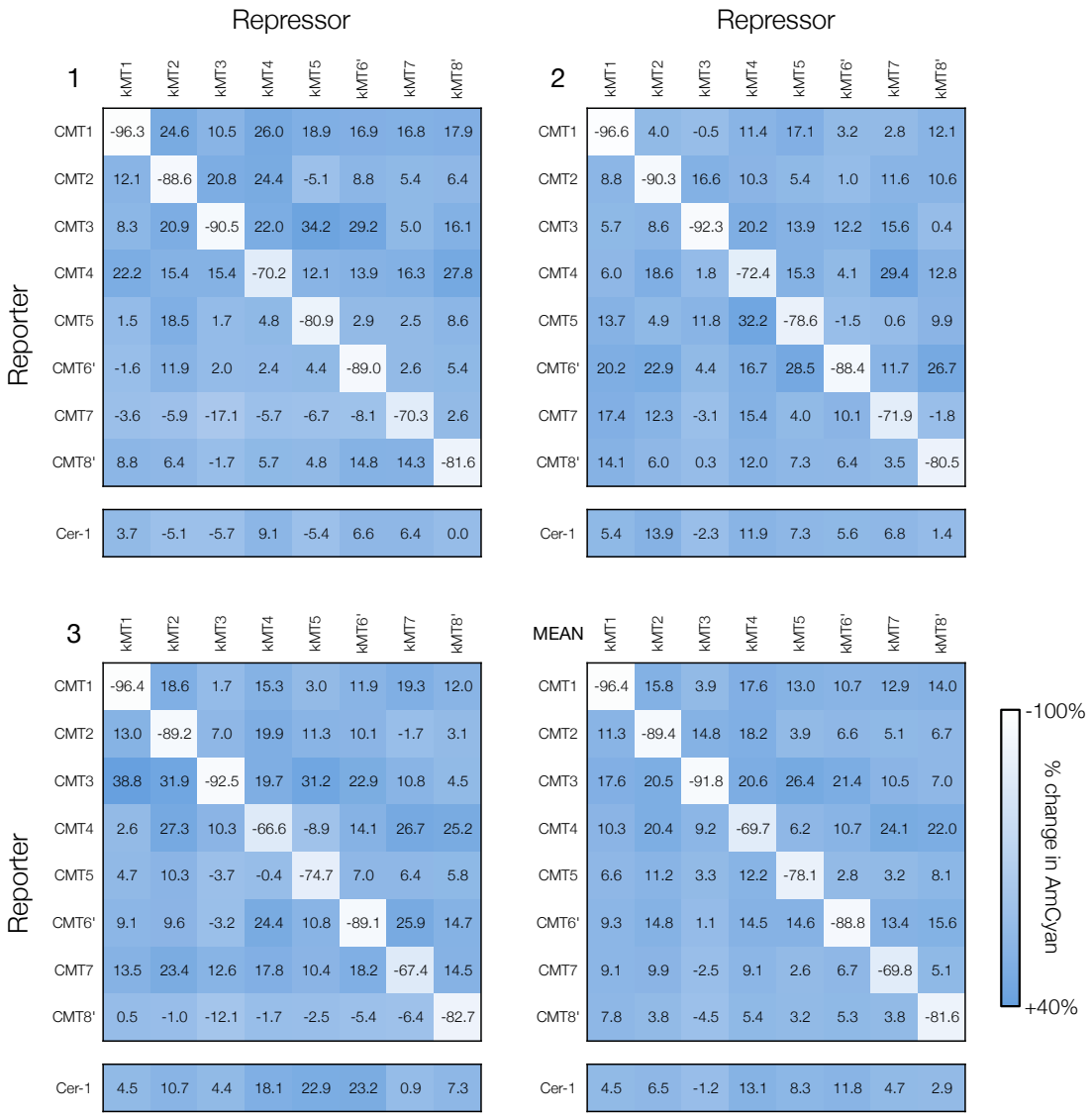

Figure 7.11: **Mean percent change.** The data is normalised against the negative control mK (see Figure 7.10) to show the percent change and averaged across the three experimental replicates.

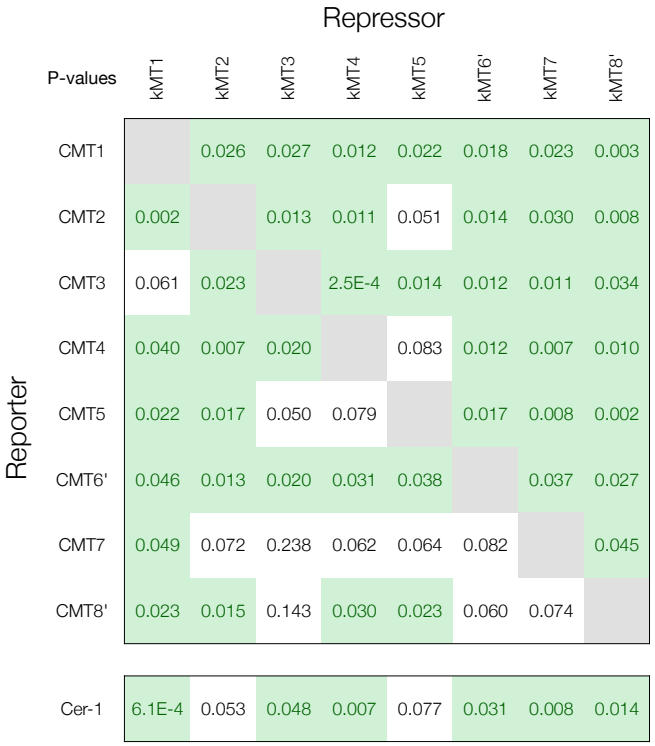

Figure 7.12: **Lack of repression in mismatched cotransfections confirmed statistically in most cases.** The figure shows P-values in a t-test that tries to reject the hypothesis that repression of at least 10% occured in a sample. The P-values <0.05 are highlighted in light green. In several cases the cutoff has not been achieved, so the null hypothesis cannot be rejected. More repeats might be needed to achieve statistically significant rejection.

Table 7.2: P-values of the ANOVA tests for samples across the miRNAs. In each case the samples of matching miRNAs and target sites have been taken out of the test in order to test for differences in the cases that were meant to be orthogonal. In the second row also the negative controls have been taken out. The results show that there is a detectable difference in means between the CMT8' samples.

	CMT1	CMT2	CMT3	CMT4	CMT5	CMT6'	CMT7	CMT8'
kMT1-8, mK	0.109	0.510	0.213	0.053	0.678	0.050	0.593	0.057
kMT1-8	0.350	0.653	0.558	0.174	0.799	0.191	0.656	0.049

groups where the target sites and miRNAs match. However, the fact that ANOVA is not directional and the samples tend to have higher means than the negative controls skews those results. Therefore, the same tests were performed, but also taking out the negative controls to see whether there are significant differences between the mismatched samples (Table 7.2, second row). The results indicate that samples differ within CMT8' ($p = 0.049$). Indeed, the kMT3-CMT8' sample has the lowest mean and taking it out further from the test increases the P-value dramatically to 0.786.

To summarise, given that the samples for the mismatching target sites and miRNAs could be expected to be higher than the negative controls, there is a detectable albeit small repression of the MT8' target site by the MT3 miRNA.

7.6 Two potentially off-targeting sequences

Section 6.3 of the experimental design chapter mentions two sequences that were produced with a high predicted off-targeting score, but scoring well against the siRNA design rules. The planned pilot experiment was aimed at testing whether such sequences would perform with low efficiency due to the dilution effect caused by the multiplicity of potential targets. The two sequences were tested in a single repeat and the results—the percent change with respect to the negative control—are shown on Figure 7.14.

Although OT2 indeed causes low repression of just above 20%, the high repression of OT1 (above 80%) lends itself to the hypothesis that an miRNA can only be diverted from its intended target by off targets that have similar or stronger

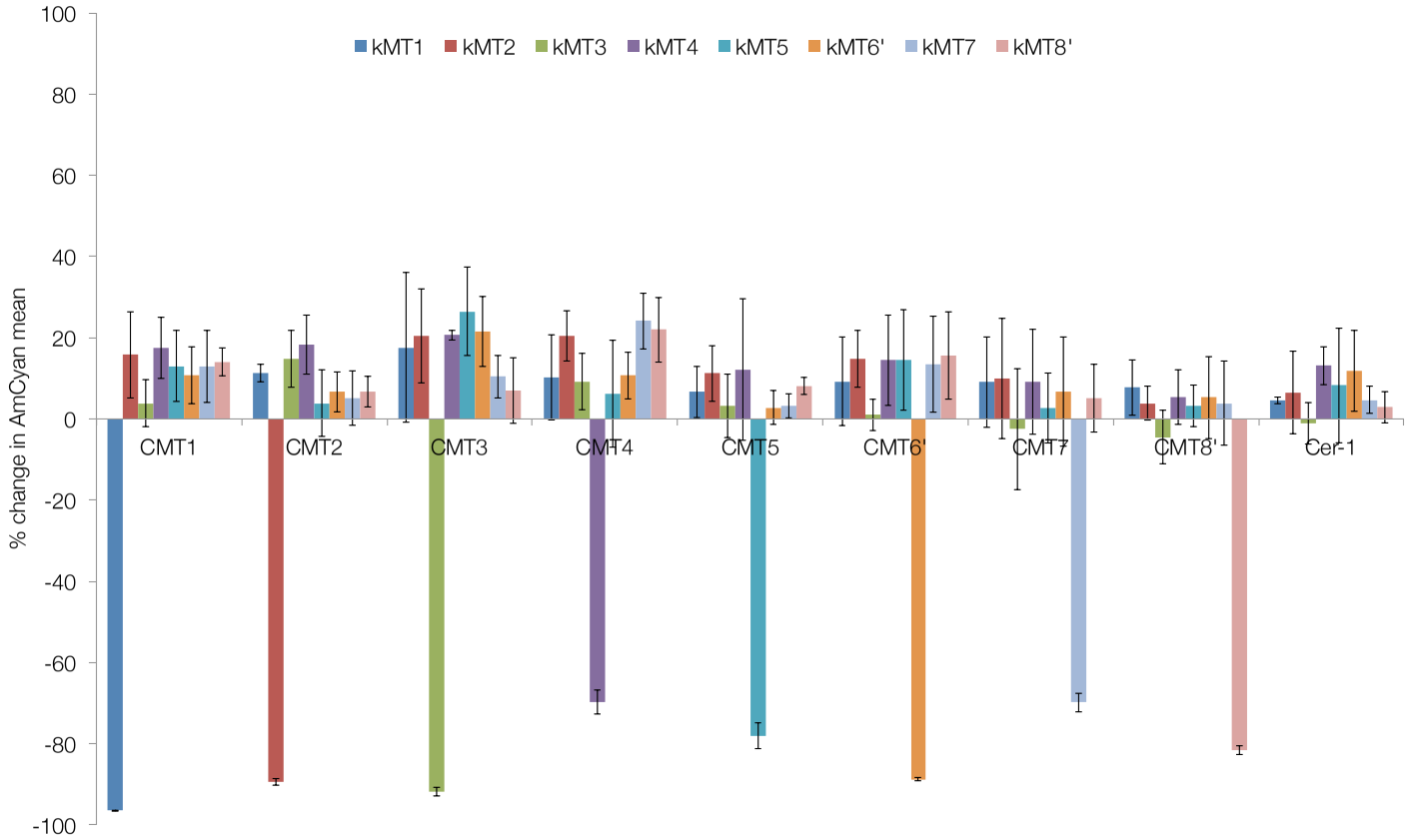

Figure 7.13: **Targets put under miRNA control show less variation than unrepressed ones.** The figure shows the same data as the bottom right heat map on Figure 7.11, but with error bars representing the standard deviation between the experimental repeats. The graph suggests a tendency of an increased expression of Cerulean in the mismatched samples. This can be explained by the fact that the hEF1a promoter used in all constructs is strong and the metabolic burden put on the cell is high in these experiments. The fact that the production of mKate with intronic miRNA is lower than mKate without the intron (see Figure 7.6) may mean that there are more cellular resources available to produce the Cerulean protein.

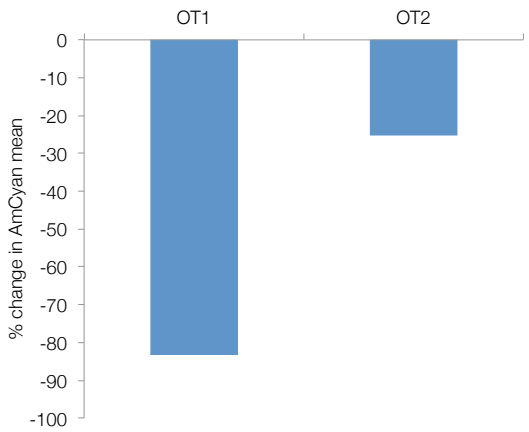

Figure 7.14: **MiRNAs bearing high off-targeting scores need not have low repression efficiency.** The figure shows a single experimental repeat of matching coexpressions of the two sequences (OT1-2) designed to have high off-targeting scores.

affinity than the intended target.

7.7 Summary

The wet lab experimental results presented in this chapter were produced with the aim of testing the efficiency and inter-ensemble orthogonality of the sequences designed in Chapter 3. Additionally, a small experiment was conducted on two sequences with predicted high off-targeting score to assess whether it would impact their efficiency.

The very first batch of the sequences failed the tests entirely due to a technical problem—the length of the miRNA sequences in conjunction with the secondary structure of the pri-miRNA of the expression template prevented their efficient processing.

A recalculated batch of eight longer sequences was tested afterwards. Two of those did not yield satisfactory efficiency in preliminary tests, but rationally chosen alternatives performed as expected.

The data also suggested no or low level (kMT3-MT8') of cross-talk between the designed miRNAs, although statistical significance has not been achieved in each case.

Finally, a small study on two highly off-targeting sequences suggested that a multitude of weakly binding targets need not divert a miRNA from its fully complementary target.

Chapter 8

Conclusion

8.1 Introduction

The study was set out to provide rational means for discovery of novel microRNA sequences. In order for synthetic biology to progress, and for the engineers to be able to build more complex synthetic systems, they need an expanded repertoire of genetic parts. One of the biggest road blocks so far has been the inability to precisely control the behaviour of the synthetic circuits (Purnick and Weiss, 2009). That is why it is important to expand the engineer's toolbox with different mechanisms of genetic regulation and more instances or regulatory elements.

MicroRNAs are short RNA sequences that translationally repress protein production (Bartel, 2004). The repression is contingent upon full or partial complementarity between the miRNA and part of the targeted mRNA sequence. A functional partial complementarity is required to have a particular structure that is roughly understood.

Much of synthetic biology today is done in prokaryotes due to their relative simplicity. Thus, microRNAs, which are only available in eukaryotes, have not received a great deal of attention. However, more complex, multicellular applications are likely to be only possible using eukaryotic cells. Also, therapeutic applications of synthetic biology must be pursued in the eukaryotic context.

The RNAi pathway which miRNAs are part of is a very robust regulation mechanism that imparts stability and allows for fine tuning of genetic circuits

(Stefani and Slack, 2008). Not many miRNAs have been carefully studied yet, but they have already been found to play diverse roles, mostly in development (Lagos-Quintana et al., 2003; Sempere et al., 2004). It could be argued that there could not be multicellular organisms without miRNAs.

MicroRNAs are very flexible from an engineering point of view. A gene can be easily put under control of an miRNA simply by tagging it with an appropriate target site in its 3' UTR. A 3' UTR can contain more than one target site, therefore allowing regulation of the strength of the repression or putting one gene under regulation by more than one miRNA.

Furthermore, miRNAs are economical—their production is cheap for the cell, which is important since one of the biggest challenges in synthetic biology is preventing excessive metabolic burden imposed on the cell by synthetic circuits. MicroRNAs also perform faster than transcription factors because they act downstream in the protein production pathway, repressing translation, while the effect of repression by transcription factors is delayed by natural mRNA degradation time.

In conjunction with the more familiar transcription factors, miRNAs provide a new dimension in the design of regulatory circuits. It is therefore only natural that libraries of robust and characterised miRNA sequences should be assembled.

In the case of transcription factors and their promoter counterparts, naturally occurring elements are picked, characterised, and used in synthetic applications (Kelly et al., 2009). However, every time a transcription factor that is endogenous to the host is used, there is a potential for cross-talk between the synthetic circuit and the background of genetic systems in the cell.

MicroRNAs are different from transcription factors in that they are short and their functional form is linear, rather than three-dimensional which makes it simpler. It is therefore much more feasible to design novel miRNA sequences and ensure their orthogonality with respect to the host system.

The design of novel microRNAs brings its own set of challenges however. An miRNA can target any part of the mRNA sequence that is not tightly folded or already occupied by any RNA-binding protein. Given that full complementarity is not required for the repression to work, this dramatically limits the 4^{22} possibility space for the novel miRNAs and makes it highly host-specific. A computational

method was needed that filteres the candidates according to potential off targeting in the transcriptome.

The technical requirements set by the miRNA biogenesis process impose further constraints. Not all of those are fully recognised and currently the best classification method for efficiently processed miRNAs is embedded in the black box machine learning methods such as support vector machines (McQuisten and Peek, 2009). Such models are trained based on data from high-throughput siRNA experiments that measure repression efficiency of different sequences. The information that comes from such experiments may mix different factors as the efficiency may be affected in many ways. Those factors will include the efficiency of biogenesis, but may also include the transcriptomic background and other, unknown factors.

The best recognised factor that affects si/miRNA efficiency is the free energy imbalance between the ends of the guide and passenger strand duplex (Reynolds et al., 2004). The imbalance can be created because some base pairs have stronger binding than others. It then decides on the rate with which each of the two are chosen as the mature miRNA.

It is also important that the novel sequences produced do not cross-talk with each other, so that they can be used in a single circuit and behave predictably.

The three requirements that stem from the analysis above—host orthogonality, efficiency, and inter-ensemble orthogonality—have been used to drive the construction of the computational process that designs synthetic miRNA sequences. The ingredients were: the host transcriptome, the siRNA design rules, RNA hybridisation simulation algorithm, and the state of the art knowledge about the structure of a working miRNA:target site pair (Brodersen and Olivier, 2009).

In order to achieve computational tractability, the designed process employed a series of heuristics in several steps and interleaved efforts to ensure the requirements at different stages.

The process was still computationally expensive and a computational cluster was used to speed up the calculations. The available cluster computing tools were either inadequate to accommodate the rapidly changing code base, or their availability was endangered due to their commercial source. For that reason, a new Java-based computing cluster framework was implemented that was efficient, flexible enough to allow for rapid code change and experimentation, while ensuring

robustness in spite of running on a cluster with failure prone nodes.

8.2 Empirical findings

The computational process yielded an ensemble of eight miRNA sequences (MT1-8). Noticeably, their off-target scores were not zero—it is not guaranteed that they will not target anything in the host genome. However, the possible range of off-targeting scores is large. The number of potential off-targets calculated for the OT1 and OT2 sequences was two orders of magnitude larger than for any of the designed MT1-8.

The two OT sequences were produced for a small study to determine whether the efficiency of a miRNA can be driven down by a high off-targeting score. This hypothesis was driven by a known effect where the repression on one target can be diluted if off-targets are present (Ebert et al., 2007). One of the OT sequences did perform efficiently however. This pointed to the importance of the affinity between the miRNA or the target. Since the gene in the test construct was tagged with a target site of full complementarity with the miRNA, the many weak off targets in the transcriptome were not able to divert the miRNA towards themselves. The OT sequences might also bind preferentially to their designed target because it is probably expressed at a much higher level than the off-targets.

This phenomenon calls for imparting a greater importance to off-targets that have a high degree of complementarity with the considered miRNA candidate.

The eight designed sequences were first tested for efficiency. Two of them—MT6 and MT8—caused the repression of their targets of only 30% or less. They were replaced with similar alternatives, but with better free energy imbalance of the guide and passenger strand duplex to aid the correct choice of the guide strand. Both of those alternatives had much better efficiency.

It is likely that the computational process did not put enough emphasis on ensuring good energy imbalance of the duplex. Especially MT8 was hindered in the early stages because its seed was allowed too many G/C nucleotides. The SVM-based siRNA design rules were not enough to ensure efficiency in these cases as all of the chosen MT sequences were highly and similarly scored. An improvement of the computational process could see a more stringent approach

in terms of ensuring the favourable free energy imbalance.

The miRNAs in the final ensemble have been characterised to cause repression between 70 and 96%. In some applications those differences may be of small importance, but the characterisation provides a guide for engineers about which miRNA to choose or a sequence that is too strong or too weak with another. The strength of the repression can be increased in each case by tagging the target with more than one copy of the target site.

In order to confirm the inter-ensemble orthogonality of the sequences, all 64 combinations of miRNAs and target sites we cotransfected and potential repression of the target measured. The results in every but one case were not indicative of repression in the mismatching pairs. However, there was relatively high variability between the repeats. Such experiments are labour intensive and only three replicates were performed. Therefore, due to insufficient amount of data, not all cases have been statistically confirmed. The most questionable is potential repression of MT7 target site by the MT3 miRNA, which had the most erratic results with one repeat showing a repression of 17% and another—an increase of expression of the target by over 12%.

A subtle effect of the MT3 miRNA on the MT8' target site was detected by the ANOVA test by comparing the mismatched samples for each experiment involving the MT8' target site. The samples in this case were not elevated above the negative control to the same extent that in the other mismatches.

8.3 Applications

This work has various implications in eukaryotic synthetic biology. The unique miRNA properties of modularity, tunability, and fast response make them a powerful addition in the synthetic biology toolbox. This project provided a source for rationally designed novel miRNAs with properties such as efficiency and orthogonality.

On the circuit level, miRNAs can be used to sharpen the expression profile both temporally and in quantitatively. The transfer curve of promoters can be altered to better match the characteristics of their input and output. For instance, response of leaky promoters can be tightened by repressing translation of the

mRNAs that are produced when the promoters are in the idealised 'off' state.

NOR gates can be easily constructed by putting one gene under control of more than one miRNA. This is done trivially by putting the relevant target sites in succession in the 3' UTR of the gene. The only technical requirement for that is to put spacing/insulating sequences between the target sites.

On the application level the most obvious area of improvement is in medical therapy. MicroRNAs are already being used in development of cancer therapies both in detection of a specific miRNA profile and in complex *in vitro* computation that classifies cancerous and healthy cells (Karagiannis and El-Osta, 2005; Xie et al., 2011).

It is this *in vitro* computational part where the artificial miRNAs are useful. MicroRNAs put into complex logic circuits can enable therapies that require intelligent actuation depending on the conditions or even timing.

8.4 Further work

There are several avenues for further work than this project could be followed with. This section relates only to the main subject of the thesis. Further work for Kaichu is presented in Chapter 4.

8.4.1 Weighting off-targets

The computational framework could be improved by providing more weight to off-targets with high level of complementarity to the candidate miRNA. Currently, all potential off-targets are treated equally, but the results of the efficiency of the OT1 and OT2 sequences suggest that weak off-targets are of lesser importance. Furthermore, it is most likely impossible to find sequences with no predicted off-targets, which suggests that their number is overestimated due to the targeting rules being too liberal at the moment. An improved version of the miRNA design process would probably use a threshold of complementarity level above which candidates would be rejected even if only one such off-target was found. The other off-targets might be weighted depending on the level of complementarity.

The understanding of what makes a working miRNA:target site pair is still

very incomplete. Any improvement in that area would allow the off-target prediction to be more accurate.

8.4.2 Ensuring favourable free energy imbalance in the duplex

A further improvement would solve the problem of insufficient free energy imbalance between the ends of the passenger and guide strand duplex. Firstly, the application of the siRNA design rules to the seed would have to be more stringent in requiring more A/T nucleotides. Furthermore, since the SVM-based siRNA rules do not enforce good enough of an imbalance, an additional filter could be applied that calculated the free energy imbalance in the candidates and rejected ones below a certain threshold.

There is an alternative to that approach however. The free energy imbalance in the duplex does not have to solely depend on the sequence of the intended guide strand. It is possible to modify the intended passenger strand to introduce mismatches and therefore bulges on the 5' end of the intended guide strand in the pre-miRNA hairpins. Indeed, natural pre-miRNA hairpins often do contain bulges (Zeng and Cullen, 2003). The bulges would go a great deal towards destabilising the preferred end of the duplex. This approach should however be carefully considered as it might introduce its own problems. Experience of this project warns against allowing bulges too close to the Drosha cleavage site, so the bulges probably could not be placed too close to the end of the duplex. Care should also be taken not to disturb the overall structure of the pri-miRNA as weakening the pre-miRNA hairpin may introduce such risk.

This approach would open up the space of possible miRNA sequences to more G/C rich ones on the seed side. It would also put using the siRNA rules in their current form in question since they are elucidated from experiments where bulge-free hairpins are used and the free energy imbalance depends only on the sequence.

Perhaps this calls also for new si/miRNA design rules that are disentangled from the duplex free energy imbalance factor.

8.4.3 More experimental repeats

However repression by the matched pairs of the miRNA and target sites has been unequivocally proven with the available data, the lack of repression in the mismatched pairs was not conclusively proven in all cases. This was due to the variability in the data, which is however typical in biological experiments. In order to deal with the uncertainty and check that the repression indeed does not happen in the mismatched cases, more experimental repeats of the cotransfection matrix would have to be performed.

8.4.4 Host orthogonality validated through proteomics

The host orthogonality requirement has not really been tested in this work due to financial constraints. The only validation in this respect was that the cells did not seem to change their phenotype dramatically or die as a result of transfection with the designed miRNAs.

One way in which host orthogonality test could be attempted could be transcriptomics. Microarray or RNA-Seq experiments could be used to assess any potential changes in the transcriptome as a result of transfection with the miRNA expression constructs.

This approach might however reveal little about the effect of the miRNAs on the cell (Selbach et al., 2008) since only targeting with high levels of complementarity results in the mRNA degradation repression mode. Only such mode could be detected with transcriptomics as only then the levels of transcripts would change.

The best way of approaching this problem would be to perform a proteome-wide measurement to assess the actual impact of the miRNAs on the cell. Procedures such as pSILAC (Selbach et al., 2008) can provide such data and should be performed once they become more accessible.

8.4.5 Characterisation of mutli-target site repression

As mentioned previously, multiple target sites tagged onto one gene increase the repression strength caused by a miRNA (Xie et al., 2011) and this effect can be

used to tune the response to the miRNA.

This work provides synthetic biologists with characterisation of repression strengths of the eight designed sequences when single target sites are used. An improvement would be to expand the reference by cases with double and quadruple target sites.

8.4.6 Investigation of promoter leakage fix with miRNA

MicroRNAs can be useful in adjusting the response of promoters. For instance, a leaky promoter that, through its transcription factor, detects a signal and then drives production of an enzyme, might produce enough enzyme to generate a response even if there is no signal present.

A potential follow up from this work might be to design, analyse and test a circuit that would utilise miRNAs that compensates for such leakage. Such design need not be trivial because the additional repression of the response signal by a miRNA might have to be disabled when the input signal arrives.

8.5 Summary

In this project a computational framework was designed and implemented for rational design of orthogonal microRNAs. Additionally, eight such sequences were produced using this method in the context of the human transcriptome. The sequences have been confirmed to be efficient experimentally. Most combinations of the miRNAs have been also shown to be orthogonal with an indication that all but one of them indeed are. These sequences are ready to use in any human-based synthetic biology application and are available physically from Gateway entry vectors.

As an offshoot of the project, an efficient and flexible cluster computing framework was implemented and released under an open-source licence.

The media attached to the text contains electronic copies of the flow cytometry data as well as images of the cells from the fluorescent microscopy for the preliminary MT1-8 results, and the three experimental replicates of the full co-transfection matrix.

Appdx A

This appendix contains a demonstration and a tutorial for a simple Hello World type application implemented in Kaichu. This application will create 10 jobs and each of the workers that will receive will sleep for the set amount of time and return a message to the master confirming its sleep time and hostname.

To start off, the parameters and the response data structures have to be defined. Both classes have to implement specific interfaces.

The class application will define the sleep time separately for each job. The `HelloParameters` class implements the `KaichuParameters` interface and contains the necessary data—sleep time in milliseconds (Listing 1).

The expected response will contain a text message, so the `HelloResponse` class, implementing the `KaichuResponse` interface contains a string. The code is shown in Listing 2.

The `HelloTask` class implements the algorithm performed by a worker for each job. The class implements the `KaichuTask` interface and overrides the `execute()` method that takes the parameters object and has to return a response object.

The task would have the worker sleep for the given time and then return a message confirming that the code run on another host. The *execute()* method follows the typical progression of reading the parameters, performing the work, and returning the results. This is shown on Listing 3.

Listing 4 ties it all together. First the `KaichuManager` object is initialised, then the task defined earlier is registered in the `TaskRegistry`, and a reference to the `DispatchQueue` is obtained. At this point the jobs can be created and enqueued for execution. The parameters state the sleep time of between 1 and 5 seconds. After all jobs are submitted Kaichu is started.

The second loop in the code deals with retrieval of the results. The `getNextResponse()`

Listing 1: The HelloParameters class defines the input for the job.

```
1 public class HelloParameters implements KaichuParameters {
2     long sleepTimeInMillis;
3     public HelloParameters(long sleepTimeInMillis) {
4         // will have the worker sleep for a given time.
5         this.sleepTimeInMillis = sleepTimeInMillis;
6     }
7 }
```

Listing 2: The HelloResponse class defines the output of the job.

```
1 public class HelloResponse implements KaichuResponse {
2     String message;
3     public HelloResponse(String message) {
4         this.message = message;
5     }
6 }
```

method on the `KaichuManager` blocks until any responses arrive from the workers and if any does, it returns them in the order of their arrival. The responses can then be unwrapped and displayed.

Listing 3: The HelloTask class implements the execute() method to define the algorithm performed as part of a job.

```
1  public class HelloTask implements KaichuTask {
2
3      @Override
4      public KaichuResponse execute(KaichuParameters params) {
5
6          // read the parameters
7          HelloParameters parameters = (HelloParameters)params;
8
9          // do the work
10         try {
11             Thread.sleep(parameters.sleepTimeInMillis);
12         } catch(InterruptedException ex) {
13             Logger.getLogger(HelloTask.class.getName()).log(Level.SEVERE, null, ex);
14         }
15
16         InetAddress address = null;
17
18         try {
19             address = InetAddress.getLocalHost();
20         } catch(UnknownHostException ex) {
21             Logger.getLogger(HelloKaichu.class.getName()).log(Level.SEVERE, null, ex);
22         }
23
24         // create and return the response object
25         return new HelloResponse("Hello! I'm " + address
26                 + " and I've just slept for "
27                 + parameters.sleepTimeInMillis + " ms!");
28     }
29 }
```

Listing 4: The main class takes care of the initialisation of the process; submission of jobs; and retrieval of results as they arrive from the cluster.

```
1  public class HelloKaichu {
2
3      final static int NUMJOBS = 10;
4
5      public static void main(String[] args) {
6          KaichuManager manager = new KaichuManager();  // create the manager
7          TaskRegistry.registerTask(new HelloTask(), manager);  // register the task
8          DispatchQueue queue = manager.getDispatchQueue(); // get the dispatch queue
9
10         // create and submit jobs
11         Random random = new Random();
12         for(int i = 0; i < NUMJOBS; i++) {
13             // have the workers sleep between 1 and 5 seconds
14             HelloParameters parameters = new HelloParameters(random.nextInt(4000) + 1000);
15
16             System.out.println("Creating a job with " + parameters.sleepTimeInMillis
17                                                 + " sleeping time.");
18
19             // create a job with: unique index, parameters, and specified task
20             KaichuJob job = new KaichuJob(parameters, HelloTask.class);
21             queue.submitJob(job);
22         }
23
24         manager.start();  // start when ready
25
26         // retrieve the responses
27         HelloResponse response = null;
28         for(int i = 0; i < NUMJOBS; i++) {
29             try {
30                 response = (HelloResponse)manager.getNextResponse();
31             } catch(InterruptedException ex) {
32                 Logger.getLogger(HelloKaichu.class.getName()).log(Level.SEVERE, null, ex);
33             }
34
35             System.out.println("Retrieved response: " + response.message);
36         }
37
38         // stop and terminate the manager
39         manager.stop();
40         manager.terminate();
41     }
42 }
```

References

Afanasjev, B., Akimov, A., Kozlov, A. and Berkovic, A. (1989), 'Graduated optimization of fractionation using a 2-component model].', *Radiobiologia, radiotherapia* **30**(2), 131. 40

Ambros, V. (2003), 'MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing', *Cell* **113**(6), 673–676. 16

Arvey, A., Larsson, E., Sander, C., Leslie, C. S. and Marks, D. S. (2010), 'Target mRNA abundance dilutes microRNA and siRNA activity', *Molecular Systems Biology* **6**(363), 1–7.
URL: *http://dx.doi.org/10.1038/msb.2010.24* 19

Bartel, D. P. (2004), 'MicroRNAs: genomics, biogenesis, mechanism, and function', *Cell* **116**(2), 281–297. 16, 17, 101

Basu, S., Gerchman, Y., Collins, C. H., Arnold, F. H. and Weiss, R. (2005), 'A synthetic multicellular system for programmed pattern formation', *Nature* **434**(7037), 1130–1134. 11

Bernard, P. and Couturier, M. (1992), 'Cell killing by the f plasmid ccdb protein involves poisoning of dna-topoisomerase ii complexes', *Journal of molecular biology* **226**(3), 735–745. 74

Bitko, V. and Barik, S. (2001), 'Phenotypic silencing of cytoplasmic genes using sequence-specific double-stranded short interfering rna and its application in the reverse genetics of wild type negative-strand rna viruses', *BMC microbiology* **1**(1), 34. 21

Blankenship, R. E., Tiede, D. M., Barber, J., Brudvig, G. W., Fleming, G., Ghirardi, M., Gunner, M. R., Junge, W., Kramer, D. M., Melis, A., Moore, T. A., Moser, C. C., Nocera, D. G., Nozik, A. J., Ort, D. R., Parson, W. W., Prince, R. C. and Sayre, R. T. (2011), 'Comparing Photosynthetic and Photovoltaic Efficiencies and Recognizing the Potential for Improvement', *Science* **332**(6031), 805–809. 10

Boutros, M. and Ahringer, J. (2008), 'The art and design of genetic screens: Rna interference', *Nature Reviews Genetics* **9**(7), 554–566. 21

Brodersen, P. and Olivier, V. (2009), 'Revisiting the principles of microRNA target recognition and mode of action', *Nature Reviews Molecular Cell Biology* **10**(FEBRUARy), 141–148.
URL: *http://www.nature.com/nrm/journal/v10/n2/abs/nrm2619.html* 19, 103

Bron, C. and Kerbosch, J. (1973), 'Algorithm 457: finding all cliques of an undirected graph', *Communications of the ACM* **16**(9), 575–577. 34

Chandran, D., Bergmann, F. T., Sauro, H. M. et al. (2009), 'TinkerCell: modular CAD tool for synthetic biology', *J Biol Eng* **3**(1), 19. 12

Chang, C.-C. and Lin, C.-J. (2011), 'Libsvm: a library for support vector machines', *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3), 27. 37

Cordes, K. R., Sheehy, N. T., White, M. P., Berry, E. C., Morton, S. U., Muth, A. N., Lee, T.-H., Miano, J. M., Ivey, K. N. and Srivastava, D. (2009), 'mir-145 and mir-143 regulate smooth muscle cell fate and plasticity', *Nature* **460**(7256), 705–710. 16

Czar, M. J., Cai, Y. and Peccoud, J. (2009), 'Writing DNA with GenoCAD™', *Nucleic acids research* **37**(suppl 2), W40–W47. 12

Davidson, B. L. and Paulson, H. L. (2004), 'Molecular medicine for the brain: silencing of disease genes with rna interference', *The Lancet Neurology* **3**(3), 145–149. 21

de Mora, K., Joshi, N., Balint, B. L., Ward, F. B., Elfick, A. and French, C. E. (2011), 'A pH-based biosensor for detection of arsenic in drinking water', *Anal Bioanal Chem* **400**(4), 1031–1039. 10

Dean, J. and Ghemawat, S. (2008), 'Mapreduce: simplified data processing on large clusters', *Communications of the ACM* **51**(1), 107–113. 51

Deans, T. L., Cantor, C. R. and Collins, J. J. (2007), 'A tunable genetic switch based on RNAi and repressor proteins for regulating gene expression in mammalian cells.', *Cell* **130**(2), 363–72.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/17662949* 11, 22

Drinnenberg, I. A., Weinberg, D. E., Xie, K. T., Mower, J. P., Wolfe, K. H., Fink, G. R. and Bartel, D. P. (2009), 'Rnai in budding yeast', *Science* **326**(5952), 544–550. 17

Ebert, M. S., Neilson, J. R. and Sharp, P. A. (2007), 'MicroRNA sponges : competitive inhibitors of small RNAs in mammalian cells', **4**(9), 721–726. 20, 22, 104

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. and Tuschl, T. (2001), 'Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells', *Nature* **411**(6836), 494–498. 16

Elbashir, S. M., Lendeckel, W. and Tuschl, T. (2001), 'RNA interference is mediated by 21-and 22-nucleotide RNAs', *Genes & development* **15**(2), 188–200. 16

Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. and Tuschl, T. (2001), 'Functional anatomy of siRNAs for mediating efficient RNAi in drosophila melanogaster embryo lysate', *The EMBO journal* **20**(23), 6877–6888. 16

Ellington, A. D. and Szostak, J. W. (1990), 'In vitro selection of rna molecules that bind specific ligands', *Nature* **346**(6287), 818–822. 14

Elowitz, M. B. and Leibler, S. (2000), 'A synthetic oscillatory network of transcriptional regulators', *Nature* **403**(6767), 335–338. 11

Fire, A., Albertson, D., Harrison, S. W. and Moerman, D. (1991), 'Production of antisense RNA leads to effective and specific inhibition of gene expression in c. elegans muscle', *Development* **113**(2), 503–514. 15

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998), 'Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans', *Nature* **391**(6669), 806–811. 15

Friedland, A. E., Lu, T. K., Wang, X., Shi, D., Church, G. and Collins, J. J. (2009), 'Synthetic gene networks that count', *Science Signaling* **324**(5931), 1199. 11

Funahashi, A., Morohashi, M., Kitano, H. and Tanimura, N. (2003), 'CellDesigner: a process diagram editor for gene-regulatory and biochemical networks', *Biosilico* **1**(5), 159–162. 12

Fung, E., Wong, W. W., Suen, J. K., Bulter, T., Lee, S.-g. and Liao, J. C. (2005), 'A synthetic gene–metabolic oscillator', *Nature* **435**(7038), 118–122. 11

Gardner, T., Cantor, C. and JJ, C. (2000), 'Construction of a genetic toggle switch in Escherichia coli.', *Nature* **403**(6767), 339–342. 11

Gashler, M., Ventura, D. and Martinez, T. (2011), 'Manifold learning by graduated optimization', *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **41**(6), 1458–1470. 40

Gilbert, D., Heiner, M., Rosser, S., Fulton, R. and Trybiło, M. (2008), 'A Case Study in Model-driven Synthetic Biology.', *BICC* **268**, 163–175. 9

Gu, X., Trybiło, M., Ramsay, S., Jensen, M., Fulton, R., Rosser, S. and Gilbert, D. (2010), 'Engineering a novel self-powering electrochemical biosensor', *Syst Synth Biol* **4**(3), 203–214. 10

Guo, S., Kemphues, K. J. et al. (1995), 'par-1, a gene required for establishing polarity in c. elegans embryos, encodes a putative ser/thr kinase that is asymmetrically distributed.', *Cell* **81**(4), 611. 15

Hamilton, A. J. and Baulcombe, D. C. (1999), 'A species of small antisense RNA in posttranscriptional gene silencing in plants', *Science* **286**(5441), 950–952. 15

Hartley, J. L., Temple, G. F. and Brasch, M. A. (2000), 'DNA cloning using in vitro site-specific recombination.', *Genome research* **10**(11), 1788–95. 73

Hobert, O. (2005), *WormBook, Specification of the nervous system*, The C. elegans Research Community.
URL: *http://www.wormbook.org* 2

Hobert, O. (2006), Architecture of a microRNA-controlled gene regulatory network that diversifies neuronal cell fates, *in* 'Cold Spring Harbor symposia on quantitative biology', Vol. 71, Cold Spring Harbor Laboratory Press, pp. 181–188. 16

Hofacker, I. L. (2003), 'Vienna RNA secondary structure server', *Nucleic acids research* **31**(13), 3429–3431. 35, 39

Holen, T., Amarzguioui, M., Wiiger, M. T., Babaie, E. and Prydz, H. (2002), 'Positional effects of short interfering rnas targeting the human coagulation trigger tissue factor', *Nucleic acids research* **30**(8), 1757–1766. 21

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. and Kummer, U. (2006), 'Copasi—a complex pathway simulator', *Bioinformatics* **22**(24), 3067–3074. 12

Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Meloon, B., Engel, S., Rosenberg, A., Cohen, D. et al. (2005), 'Design of a genome-wide siRNA library using an artificial neural network', *Nature biotechnology* **23**(8), 995–1001. 31, 36, 43, 44

Izant, J. G., Weintraub, H. et al. (1984), 'Inhibition of thymidine kinase gene expression by anti-sense RNA: a molecular approach to genetic analysis.', *Cell* **36**(4), 1007. 15

Karagiannis, T. C. and El-Osta, A. (2005), 'Rna interference and potential therapeutic applications of short interfering rnas', *Cancer gene therapy* **12**(10), 787–795. 21, 106

Kelly, J. R., Rubin, A. J., Davis, J. H., Ajo-Franklin, C. M., Cumbers, J., Czar, M. J., de Mora, K., Glieberman, A. L., Monie, D. D. and Endy, D. (2009), 'Measuring the activity of BioBrick promoters using an in vivo reference standard.', *Journal of biological engineering* **3**, 4. 8, 102

Kemmer, C., Fluri, D. A., Witschi, U., Passeraub, A., Gutzwiller, A. and Fussenegger, M. (2011), 'A designer network coordinating bovine artificial insemination by ovulation-triggered release of implanted sperms', *J Control Release* **150**(1), 23–29. 10

Kramer, B. P., Viretta, A. U., Daoud-El-Baba, M., Aubel, D., Weber, W. and Fussenegger, M. (2004), 'An engineered epigenetic transgene switch in mammalian cells', *Nat. Biotechnol.* **22**(7), 867–870. 11

Lagos-Quintana, M., RAUHUT, R., MEYER, J., BORKHARDT, A. and TUSCHL, T. (2003), 'New microRNAs from mouse and human', *RNA* **9**(2), 175–179. 16, 102

Lai, E. C. (2002), 'Micro RNAs are complementary to 3[prime] RNA sequence motifs that mediate negative post-transcriptional regulation', *Nat Genet* **30**(4), 363–364.
URL: *http://dx.doi.org/10.1038/ng865* 19

Lai, E. C. (2003), 'microRNAs: runts of the genome assert themselves', *Current Biology* **13**(23), R925–R936. 16

Lee, R. C., Feinbaum, R. and Ambros, V. (1993), 'The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.', *Cell* **75**, 843–854. 19

Lewis, B. P., Burge, C. B. and Bartel, D. P. (2005), 'Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets', *Cell* **120**(1), 15 – 20.
URL: *http://www.sciencedirect.com/science/article/pii/S0092867404012607* 19

Litzkow, M. J., Livny, M. and Mutka, M. W. (1988), Condor-a hunter of idle workstations, *in* 'Distributed Computing Systems, 1988., 8th International Conference on', IEEE, pp. 104–111. 49

Liu, Q., Qi, X. and Fan, S. (2010), 'Simulating bioreaction processes based on simbiology', *Jisuanji Yingyong yu Ruanjian* **27**(8), 212–214. 12

Lu, P. Y., Xie, F., Woodle, M. et al. (2003), 'sirna-mediated antitumorigenesis for drug target validation and therapeutics.', *Current opinion in molecular therapeutics* **5**(3), 225. 21

MacArthur, B. D., Ma'ayan, A. and Lemischka, I. R. (2009), 'Systems biology of stem cell fate and cellular reprogramming', *Nature Reviews Molecular Cell Biology* **10**(10), 672–681. 12

Malo, G. D., Pouwels, L. J., Wang, M., Weichsel, A., Montfort, W. R., Rizzo, M. A., Piston, D. W. and Wachter, R. M. (2007), 'X-ray structure of cerulean gfp: a tryptophan-based chromophore useful for fluorescence lifetime imaging', *Biochemistry* **46**(35), 9865–9873. 73

McCaffrey, A. P., Nakai, H., Pandey, K., Huang, Z., Salazar, F. H., Xu, H., Wieland, S. F., Marion, P. L. and Kay, M. A. (2003), 'Inhibition of hepatitis b virus in mice by rna interference', *Nature biotechnology* **21**(6), 639–644. 21

McQuisten, K. A. and Peek, A. S. (2009), 'Comparing Artificial Neural Networks, General Linear Models and Support Vector Machines in Building Predictive Models for Small Interfering RNAs.', *Plos One* **4**(10), e7522. 20, 36, 37, 103

Miska, E. A., Alvarez-Saavedra, E., Abbott, A. L., Lau, N. C., Hellman, A. B., McGonagle, S. M., Bartel, D. P., Ambros, V. R. and Horvitz, H. R. (2007), 'Most caenorhabditis elegans microRNAs are individually not essential for development or viability', *PLoS Genet* **3**, e215. 3

Molnár, A., Schwach, F., Studholme, D. J., Thuenemann, E. C. and Baulcombe, D. C. (2007), 'miRNAs control gene expression in the single-cell alga chlamydomonas reinhardtii', *Nature* **447**(7148), 1126–1129. 17

Nellen, W. and Lichtenstein, C. (1993), 'What makes an mRNA anti-sensi-itive?', *Trends in biochemical sciences* **18**(11), 419–423. 15

Odersky, M., Altherr, P., Cremet, V., Emir, B., Micheloud, S., Mihaylov, N., Schinz, M., Stenman, E. and Zenger, M. (2004), 'The scala language specification'. 64

Oracle (n.d.), Java Remote Method Invocation - Distributed Computing for Java, Technical report, Oracle.
URL: *http://www.oracle.com/technetwork/java/javase/tech/index-jsp-138781.html* 53

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999), 'The pagerank citation ranking: bringing order to the web.'. 51

Pillai, R. S., Bhattacharyya, S. N. and Filipowicz, W. (2007), 'Repression of protein synthesis by miRNAs: how many mechanisms?', *Trends in cell biology* **17**(3), 118–126. 19

Purnick, P. E. and Weiss, R. (2009), 'The second wave of synthetic biology: from modules to systems', *Nat. Rev. Mol. Cell Biol.* **10**(6), 410–422. 11, 101

Reich, S. J., Fosnot, J., Kuroki, A., Tang, W., Yang, X., Maguire, A. M., Bennett, J., Tolentino, M. J. et al. (2003), 'Small interfering rna (sirna) targeting vegf effectively inhibits ocular neovascularization in a mouse model', *Mol Vis* **9**(5), 210–216. 21

Reid, T., Warren, R. and Kirn, D. (2002), 'Intravascular adenoviral agents in cancer patients: lessons from clinical trials', *Cancer gene therapy* **9**(12), 979–986. 21

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. and Ruvkun, G. (2000), 'The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans', *Nature* **403**(6772), 901–906.
URL: *http://dx.doi.org/10.1038/35002607* 3, 19

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. S. and Khvorova, A. (2004), 'Rational siRNA design for RNA interference.', *Nature biotechnology* **22**(3), 326–30.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/14758366* 17, 20, 103

Rinaudo, K., Bleris, L., Maddamsetti, R., Subramanian, S., Weiss, R. and Benenson, Y. (2007*a*), 'A universal RNAi-based logic evaluator that operates in mammalian cells', *Nat. Biotechnol.* **25**(7), 795–801. 11, 23

Rinaudo, K., Bleris, L., Maddamsetti, R., Subramanian, S., Weiss, R. and Benenson, Y. (2007*b*), 'A universal rnai-based logic evaluator that operates in mammalian cells', *Nature biotechnology* **25**(7), 795–801. 76

Rodrigo, G., Landrain, T. E. and Jaramillo, A. (2012), 'De novo automated design of small rna circuits for engineering synthetic riboregulation in living cells', *Proceedings of the National Academy of Sciences* **109**(38), 15271–15276. 13, 14

Rodrigo, G., Landrain, T. E., Majer, E., Daròs, J.-A. and Jaramillo, A. (2013), 'Full design automation of multi-state rna devices to program gene expression using energy-based optimization', *PLoS computational biology* **9**(8), e1003172. 13

Rosenfeld, A. et al. (1984), *Multiresolution image processing and analysis*, Vol. 12, Springer-Verlag New York:. 40

Saeidi, N., Wong, C. K., Lo, T. M., Nguyen, H. X., Ling, H., Leong, S. S., Poh, C. L. and Chang, M. W. (2011), 'Engineering microbes to sense and eradicate Pseudomonas aeruginosa, a human pathogen', *Mol. Syst. Biol.* **7**, 521. 10

Scherr, M., Battmer, K., Winkler, T., Heidenreich, O., Ganser, A. and Eder, M. (2003), 'Specific inhibition of bcr-abl gene expression by small interfering rna', *Blood* **101**(4), 1566–1569. 21

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008), 'Widespread changes in protein synthesis induced by micrornas', *Nature* **455**(7209), 58–63. 72, 108

Sempere, L. F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., Ambros, V. et al. (2004), 'Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation', *Genome Biol* **5**(3), R13. 16, 102

Shcherbo, D., Merzlyak, E. M., Chepurnykh, T. V., Fradkov, A. F., Ermakova, G. V., Solovieva, E. A., Lukyanov, K. A., Bogdanova, E. A., Zaraisky, A. G., Lukyanov, S. et al. (2007), 'Bright far-red fluorescent protein for whole-body imaging', *Nature methods* **4**(9), 741–746. 73

Shvachko, K., Kuang, H., Radia, S. and Chansler, R. (2010), The hadoop distributed file system, *in* 'Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on', IEEE, pp. 1–10. 51

Sioud, M. (2004), 'Therapeutic sirnas', *Trends in pharmacological sciences* **25**(1), 22–28. 21

Smith, T. and Waterman, M. (1984), 'Identification of common molecular subsequences', *J Mol Biol* **147**(1), 195–197. 33, 40

Sohka, T., Heins, R. A., Phelan, R. M., Greisler, J. M., Townsend, C. A. and Ostermeier, M. (2009), 'An externally tunable bacterial band-pass filter', *Proc. Natl. Acad. Sci. U.S.A.* **106**(25), 10135–10140. 11

Song, E., Lee, S.-K., Dykxhoorn, D. M., Novina, C., Zhang, D., Crawford, K., Cerny, J., Sharp, P. A., Lieberman, J., Manjunath, N. et al. (2003), 'Sustained small interfering rna-mediated human immunodeficiency virus type 1 inhibition in primary macrophages', *Journal of virology* **77**(13), 7174–7181. 21

Song, E., Lee, S.-K., Wang, J., Ince, N., Ouyang, N., Min, J., Chen, J., Shankar, P. and Lieberman, J. (2003), 'Rna interference targeting fas protects mice from fulminant hepatitis', *Nature medicine* **9**(3), 347–351. 21

Stefani, G. and Slack, F. J. (2008), 'Small non-coding RNAs in animal development', *Nature Reviews Molecular Cell Biology* **9**(3), 219–230. 17, 102

Stricker, J., Cookson, S., Bennett, M. R., Mather, W. H., Tsimring, L. S. and Hasty, J. (2008), 'A fast, robust and tunable synthetic gene oscillator', *Nature* **456**(7221), 516–519. 11

Sulston, J. and Horvitz, H. (1977), 'Post-embryonic cell lineages of the nematode, caenorhabditis elegans', *Developmental Biology* **56**(1), 110 – 156.
URL: *http://www.sciencedirect.com/science/article/pii/0012160677901580* 2

Tafer, H., Ameres, S. L., Obernosterer, G., Gebeshuber, C. a., Schroeder, R., Martinez, J. and Hofacker, I. L. (2008), 'The impact of target site accessibility on the design of effective siRNAs.', *Nature biotechnology* **26**(5), 578–83.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/18438400* 20

Tan, C., Marguet, P. and You, L. (2009), 'Emergent bistability by a growth-modulating positive feedback circuit.', *Nature chemical biology* **5**(11), 842–8. 27

Tigges, M., Marquez-Lago, T. T., Stelling, J. and Fussenegger, M. (2009), 'A tunable synthetic mammalian oscillator', *Nature* **457**(7227), 309–312. 11

Tuerk, C. and Gold, L. (1990), 'Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase', *Science* **249**(4968), 505–510. 14

Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R. and Saigo, K. (2004), 'Guidelines for the selection of highly effective sirna sequences for mammalian and chick rna interference', *Nucleic Acids Research* **32**(3), 936–948. 17

van Rooij, E., Sutherland, L. B., Qi, X., Richardson, J. A., Hill, J. and Olson, E. N. (2007), 'Control of stress-dependent cardiac growth and gene expression by a microRNA', *Science Signalling* **316**(5824), 575. 16

Wall, N. R., Shi, Y. et al. (2003), 'Small rna: can rna interference be exploited for therapy?', *Lancet* **362**(9393), 1401–1403. 21

Widmaier, D. M., Tullman-Ercek, D., Mirsky, E. A., Hill, R., Govindarajan, S., Minshull, J. and Voigt, C. A. (2009), 'Engineering the Salmonella type III secretion system to export spider silk monomers', *Mol. Syst. Biol.* **5**, 309. 10

Wightman, B., Ha, I. and Ruvkun, G. (1993), 'Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans.', *Cell* **75**(5), 855–62.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/8252622* 3

Win, M. N. and Smolke, C. D. (2007), 'A modular and extensible rna-based gene-regulatory platform for engineering cellular function', *Proceedings of the National Academy of Sciences* **104**(36), 14283–14288. 14

Win, M. N. and Smolke, C. D. (2008*a*), 'Higher-order cellular information processing with synthetic RNA devices', *Science* **322**(5900), 456–460. 11

Win, M. N. and Smolke, C. D. (2008*b*), 'Higher-order cellular information processing with synthetic rna devices', *Science* **322**(5900), 456–460. 14

Wohlbold, L., van der Kuip, H., Miething, C., Vornlocher, H.-P., Knabbe, C., Duyster, J. and Aulitzky, W. E. (2003), 'Inhibition of bcr-abl gene expression by small interfering rna sensitizes for imatinib mesylate (sti571)', *Blood* **102**(6), 2236–2239. 21

Wu, T., Wu, S., Yin, Q., Dai, H., Li, S., Dong, F., Chen, B. and Fang, H. (2011), '[Biosynthesis of amorpha-4,11-diene, a precursor of the antimalarial agent artemisinin, in Escherichia coli through introducing mevalonate pathway]', *Sheng Wu Gong Cheng Xue Bao* **27**(7), 1040–1048. 9

Xia, H., Mao, Q., Paulson, H. L. and Davidson, B. L. (2002), 'sirna-mediated gene silencing in vitro and in vivo', *Nature biotechnology* **20**(10), 1006–1010. 21

Xie, Z., Wróblewska, L., Prochazka, L., Weiss, R. and Benenson, Y. (2011), 'Multi-input RNAi-based logic circuit for identification of specific cancer cells', *Science* **333**(6047), 1307–1311. 10, 23, 76, 106, 108

Ye, M., Haralick, R. M. and Shapiro, L. G. (2003), 'Estimating piecewise-smooth optical flow with global matching and graduated optimization', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25**(12), 1625–1630. 40

Zaree Mahmodabady, A., Javadi, H. R., Kamali, M., Najafi, A. and Hojati, Z. (2010), 'Bcr-abl silencing by specific small-interference rna expression vector as a potential treatment for chronic myeloid leukemia', *Iranian Biomedical Journal* **14**(1), 1–8. 21

Zender, L., Hütker, S., Liedtke, C., Tillmann, H. L., Zender, S., Mundt, B., Waltemathe, M., Gösling, T., Flemming, P., Malek, N. P. et al. (2003), 'Caspase 8 small interfering rna prevents acute liver failure in mice', *Proceedings of the National Academy of Sciences* **100**(13), 7797–7802. 21

Zeng, Y. and Cullen, B. R. (2003), 'Sequence requirements for micro rna processing and function in human cells', *Rna* **9**(1), 112–123. 43, 82, 107

Zuker, M. and Stiegler, P. (1981), 'Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information', *Nucleic acids research* **9**(1), 133–148. 40